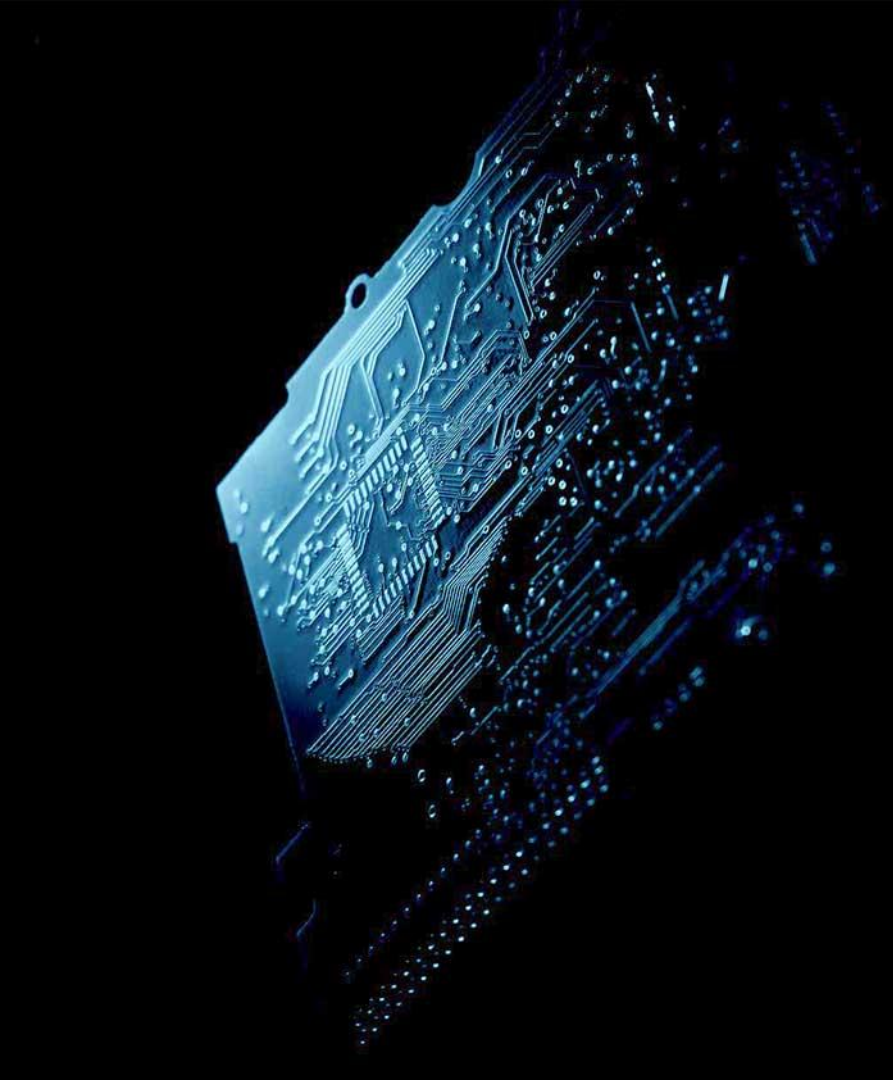
Infrastructure for the AI Era:

Enabling Enterprises to Upgrade from Hardware and Software to Architecture and AI Applications

Speaker

Fred Jhang, Lead of CE Front End Engineering, GMI Cloud



We all know AI is important,
but the real question is —

how do we actually start?

Who we are:

Introducing GMI Cloud

One out of 6 Nvidia Reference Platform Cloud Partner

APAC Focused

AI native GPU Platform

Full-stack Cluster Engine Inference Engine

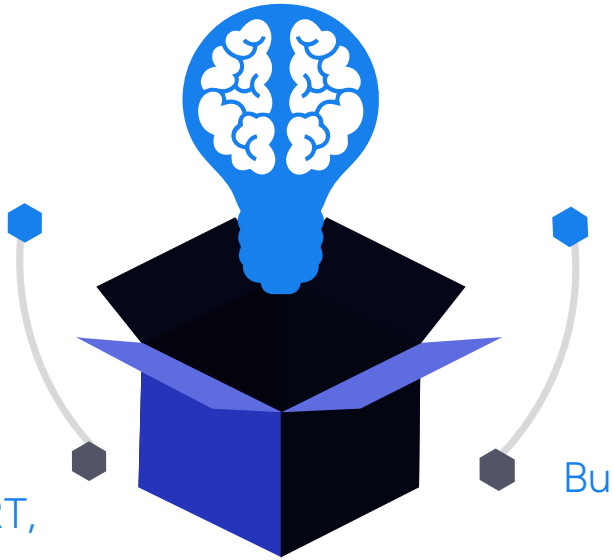
NVIDIA Preferred Partner

Level 1:

Hardware Layer

Optimized GPU servers with CUDA, NVLink, and InfiniBand

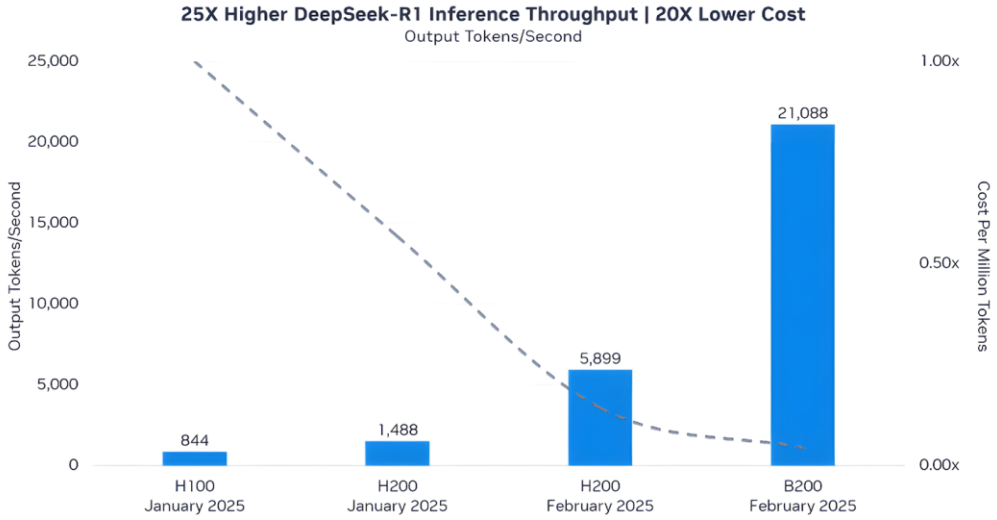
Optional pre-installed AI environments (PyTorch, TensorRT, vLLM, Triton, etc.)



Optional setup for Kubernetes or OpenStack environments, enabling containerized and scalable AI workloads

Built-in monitoring agents powered by DCGM and Prometheus

Always on the Lookout for Best Hardware



Throughput tested on single 8 x H100、H200、B200 server node

*Data provided by NVIDIA DeepSeek-FP4

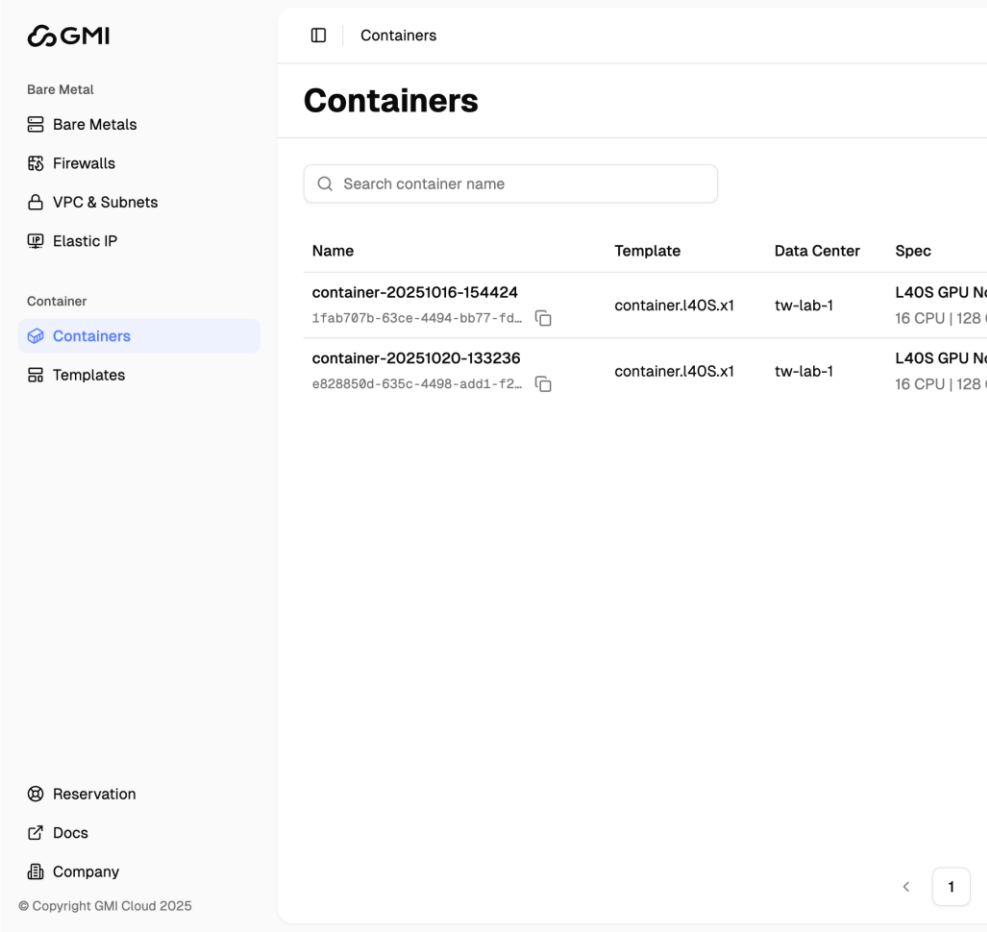
B200 is more than 25x performant than H200. Latest hardware hosts latest model better.

Hardware upgrades is one of the easiest way to improve performance.

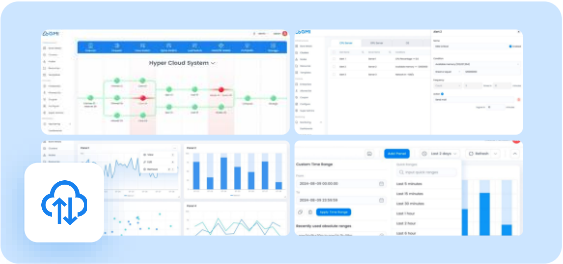
Level 2:

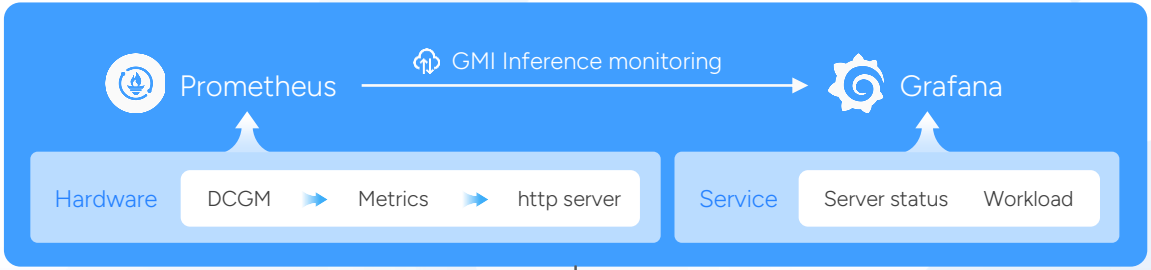
Management Layer

- **Cluster Engine** provides both Web and API control interfaces
- One-click creation of Bare Metal, Container, or Kubernetes Clusters
- Integrated monitoring and health checks



Global End to End Monitoring





K8s Cluster

K8s control plane

K8s control plane

K8s control plane

K8s Worker

GPU GPU GPU

GPU

Memory

Disk

K8s Worker

GPU GPU GPU

GPU

Memory

Disk

Storage

K8s Cluster

K8s control plane

K8s control plane

K8s control plane

K8s control plane

K8s Cluster

K8s control plane

K8s control plane

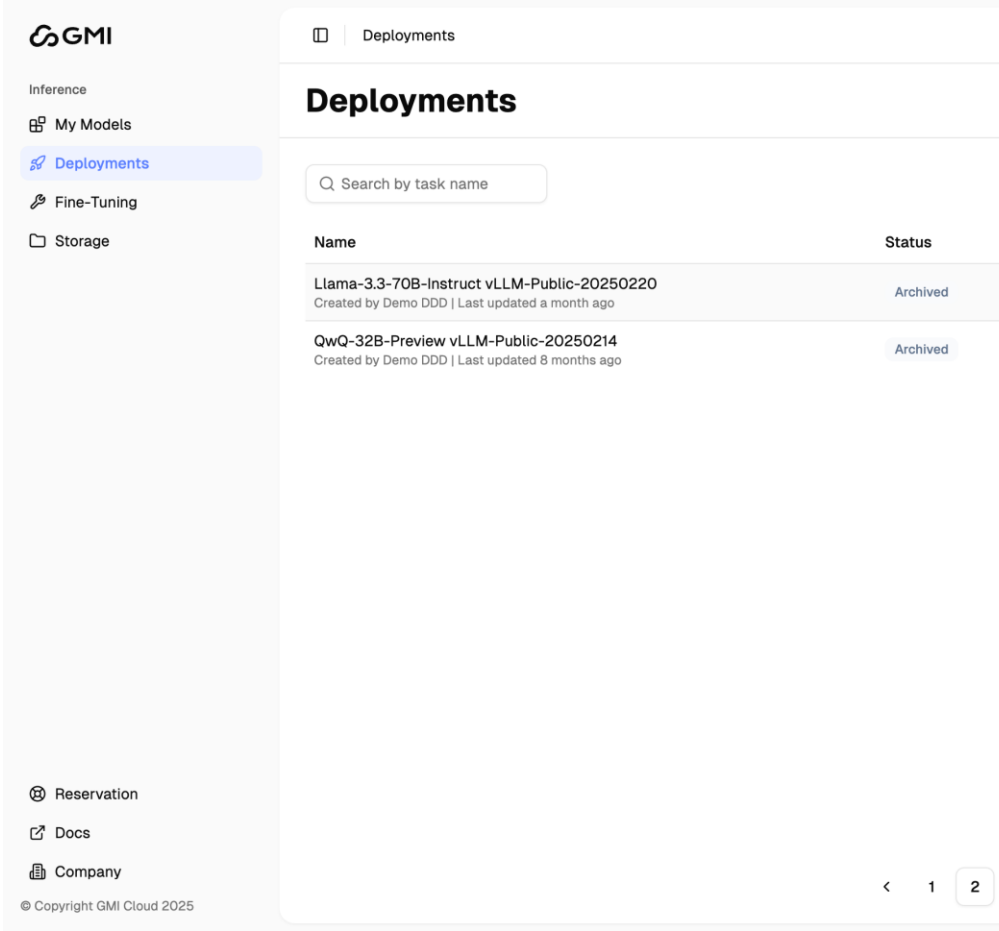
K8s control plane

K8s control plane

Level 3:

Model Serving Layer

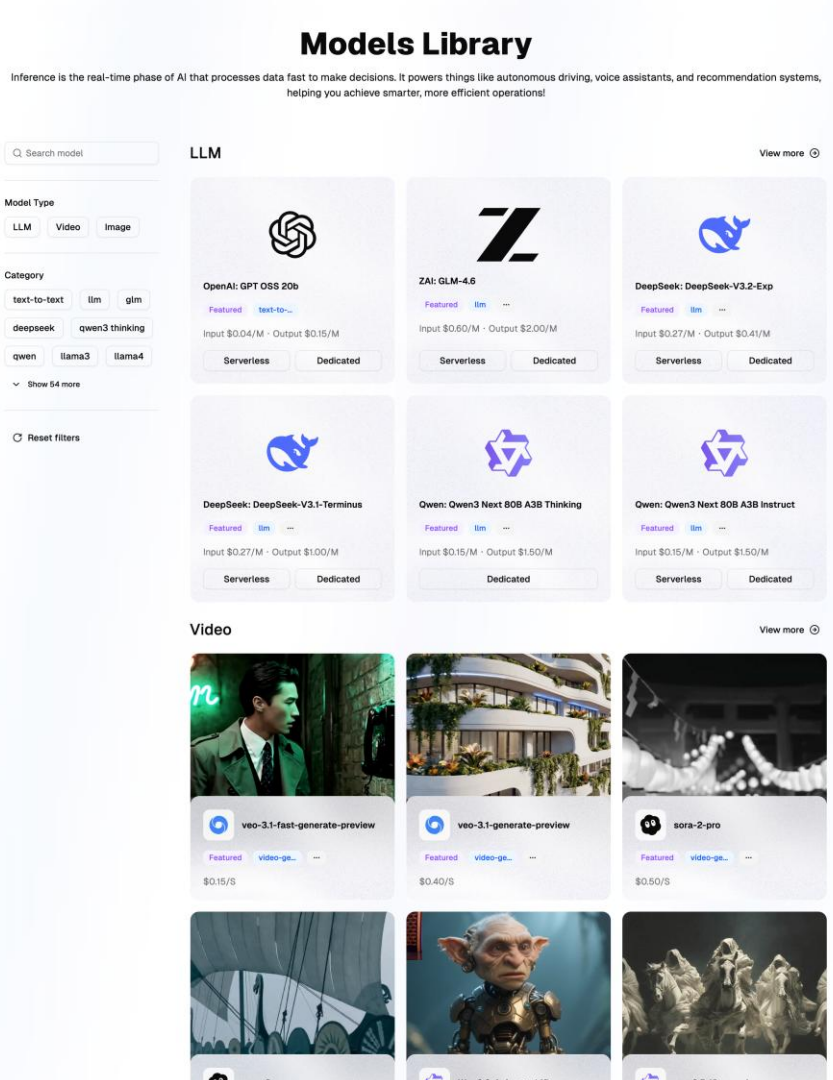
- **Inference Engine** provides a unified platform for model deployment
- One-click deployment of vLLM / SGLang inference engines
- Supports Auto Scaling, OpenAI-compatible APIs, and load balancing



Level 4:

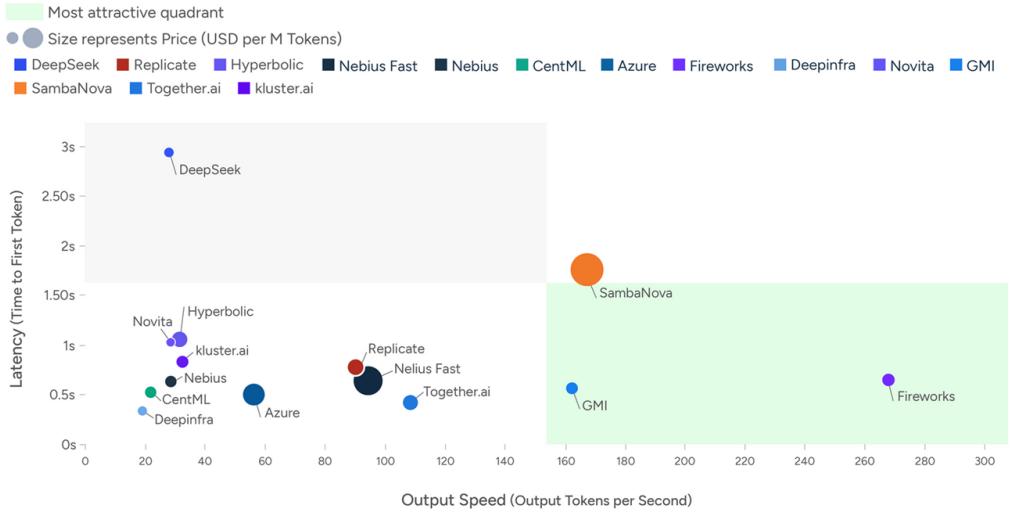
Optimized Model Layer

- **Inference Engine** offers a multimodal inference platform supporting text, image, and video generation
- Integrates and optimizes open-source models such as Llama, DeepSeek, Mistral, Sora, and Flux
- Supports quantization, multi-node inference, and cross-region caching with TeaCache
- Provides ready-to-use APIs, enabling rapid integration into enterprise AI applications



Who we work with on Inferencing

Industry-Leading Inference Performance



Latency vs. Output Speed:

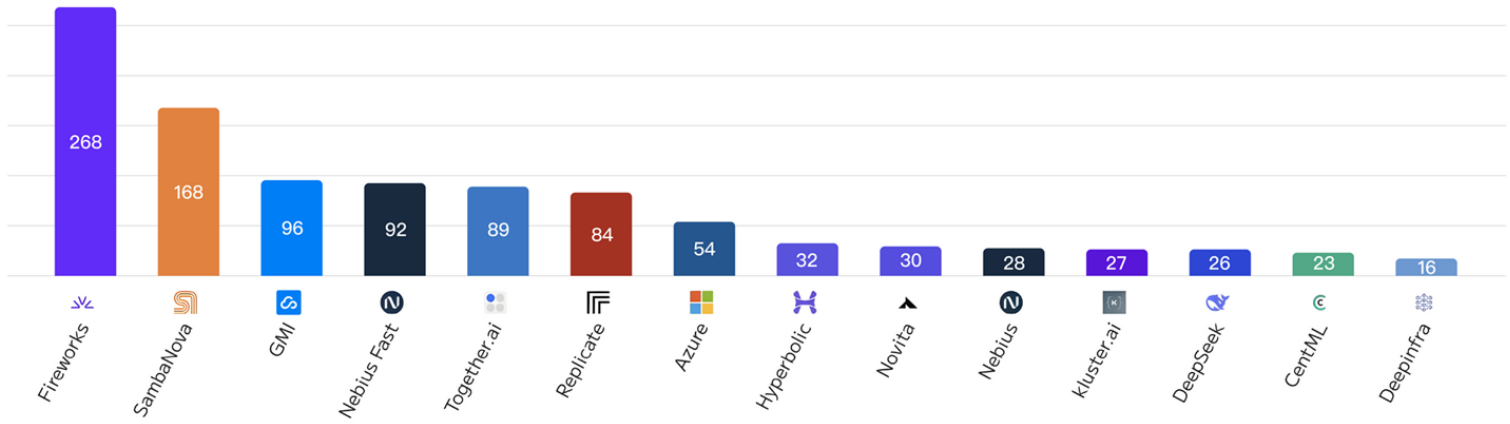
DeepSeek V3 0324 (Mar'25) Providers

Latency: Seconds to First Token Received; Ouput Speed: Ouput Tokens per Second; 1,000 Input Tokens

Industry-Leading Inference Performance

Output Speed

Output Tokens per Second, Higher is better



The Value of Efficient Inference (same hardware)

30x

Performance boost

Due to bottlenecks in model execution

71%

Cost Reduction

Possible with proper optimization

11x

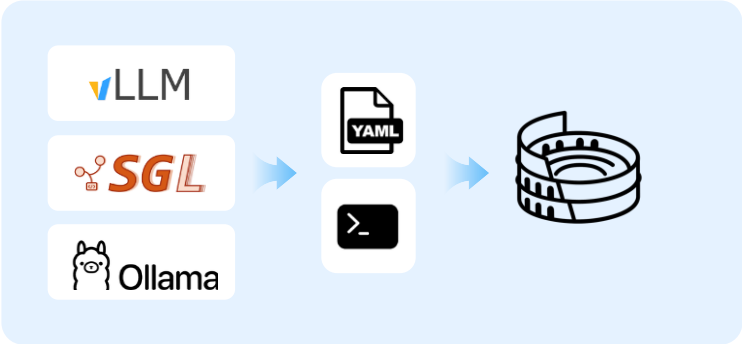
Inference speed

Due to challenges balancing latency and throughput

Introducing GMI Inference Engine Benchmark tool

Open Source Community Release

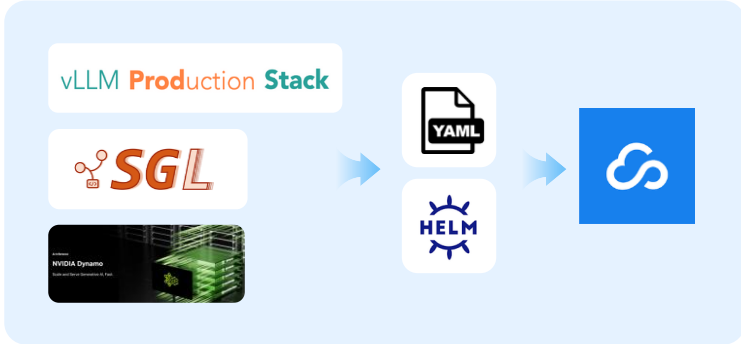
A framework for evaluating various large model inference engines (e.g., vLLM, SGLang, Ollama).



Deployment utilizing a single machine command line

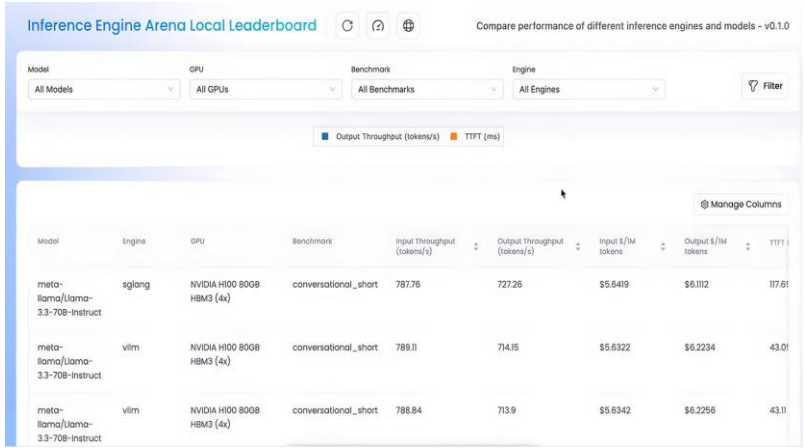
GMI Cloud Edition

A service that facilitates benchmarking and assessment of diverse production-grade inference scenarios, including vLLM, SGLang, the vllm production stack, and Dynamo.



Online cluster deployment to guarantee optimal hardware and network performance

GMI Inference Engine Benchmark tool



Initiate a dedicated inference endpoint utilizing artifact xyz...

Benchmark

GMI

Conduct benchmarks and incorporate new records.

Engage with the core of the GMI Cloud platform:

1. The inference configuration can be initiated or replicated with a single click via the GMI Cloud platform.
2. If data for the necessary configuration is not yet available, new benchmarks can be executed immediately via the GMI Cloud platform.

GMI

Thank you

Welcome to GMI Cloud

278 Castro St, Mountain View, CA 94041

gmicloud.ai