MangoBoost Sets the Highest MLPerf Inference v5.0 Result in History for Llama2-70B-Offline

Achieved with MangoBoost LLMBoost™ AI Enterprise MLOps Software
that scales to Multi-Node Servers powered by AMD Instinct™ MI300X GPUs

MANGOBOOST

REVISION 1.0 | Apr 3, 2025

Disclaimer

The performance claims in this document are based on the internal cluster environment. Actual performance may vary depending on the server configuration. Software and workloads used in performance tests may have been optimized for performance only on MangoBoost products. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. Results that are based on pre-production systems and components as well as results that have been estimated or simulated using MangoBoost reference platform for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. MangoBoost does not guarantee any specific outcome. Nothing contained herein is, or shall be relied upon as, a promise or representation or warranty as to future performance of MangoBoost or any MangoBoost product. The information contained herein shall not be deemed to expand in any way the scope or effect of any representations or warranties contained in the definitive agreement for MangoBoost products.

The information contained herein may not be reproduced in whole or in part without prior written consent of MangoBoost. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. MangoBoost assumes no obligation to update or otherwise correct or revise this information and MangoBoost reserves the right to make changes to the content hereof from time to time without any notice. Nothing contained herein is intended by MangoBoost, nor should it be relied upon, as a promise or a representation as to the future.

MANGOBOOST MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

Table of Contents

01 | Key Takeaways

MangoBoost demonstrates **LLMBoost™ AI Enterprise MLOps software** in its ground-breaking MLPerf Inference v5.0 submission on **AMD MI300X GPU servers**

First-ever multi-node MLPerf Inference result on **AMD Instinct MI300X GPUs**

Sets **highest record for Llama2-70B** (offline), beating all other MLPerf Inference results to date

24% higher performance vs. 32x H100 GPUs published MLPerf result

Lower hardware cost, better efficiency, no compromise

One-line deployment with OpenAI-compatible APIs

Supports **50+ open models**, hardware-flexible, deployable anywhere (cloud/on-prem)

Beyond LLMBoost software, MangoBoost also offers **DPU hardware acceleration solutions** for AI and cloud infrastructure

MangoBoost is proud to announce a ground-breaking **MLPerf inference v5.0** submission: we are the **first-ever multi-node MLPerf submission using AMD Instinct™ GPUs**, achieving the **highest result in Llama2-70B offline inference category in history, beating all other submitted results.** By scaling across **32 MI300X GPUs over four server nodes**, MangoBoost has not only showcased the scalability of **AMD Instinct for large-scale AI workloads**, but also the **robustness and high performance of our MLOps enterprise software - LLMBoost.**

02 | Introduction

LLMBoost: Easy to Use, High Performance, Scalable, and GPU-flexible

LLMBoost is an easy-to-use AI Enterprise MLOps software that offers flexibility and performance. It runs seamlessly on both AMD and NVIDIA GPUs, and supports open 50+ models like LLaMA, Qwen, DeepSeek, and multimodal models like Llava, with no custom orchestration required. Summary of LLMBoost highlights are as follows:

- **One-line deployment** via Docker
- **Built-in OpenAI-compatible and REST APIs**
- **Cloud-ready** (available on Azure, AWS, and GCP)
- **On-premises friendly** for full control, security, and performance
- **Breakthrough performance** – Up to 138× faster than Ollama and 74× faster than HuggingFace
- **Unmatched cost-efficiency** – Up to **99% lower GPU cost per million tokens** compared to Ollama, and **significantly more efficient than vLLM and HuggingFace TGI**

LLMBoost integrates several patent-pending technologies that further enhance the performance and scalability, including:

- **Auto Parallelization** – Efficiently distributes large models across GPUs and nodes
- **Auto Config Tuning** – Optimizes runtime parameters based on workload characteristics
- **Auto Context Scaling** – Dynamically adapts LLM's memory usage to maximize GPU utilization
- **Auto Disaggregated Deployment** – Flexible deployment across multiple inference stages

No vendor lock-in. No complex setups. Just performance, efficiency, and simplicity.

03 | MangoBoost MLPerf Inference Result

Powered by AMD MI300X, Supercharged by LLMBoost Software

MangoBoost record-breaking MLPerf performance was made possible through deep collaboration with AMD and full integration with the AMD ROCm software stack, unlocking the full potential of MI300X GPUs with industry-leading compute density and massive memory bandwidth.

Together, AMD's ROCm platform and MangoBoost's LLMBoost stack deliver an AI inference solution that is fast, scalable, and easy to deploy—whether on a single node or across a multi-node cluster.

LLMBoost on 32x MI300s Outperforms 32x H100s – At Lower Cost

In this MLPerf submission, LLMBoost delivers a 24% higher throughput than the previous best multi-node MLPerf result from Juniper Networks using 32 NVIDIA H100 GPUs[1], which reported 82,749 tokens per second (TPS) on Llama2-70B offline inference.

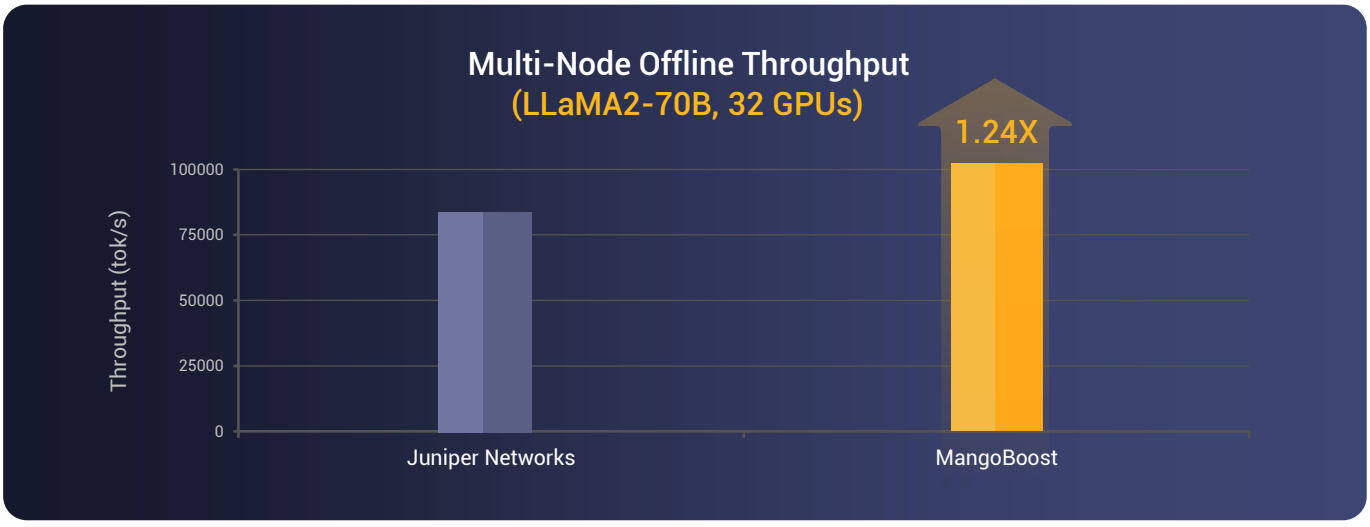

FIGURE 1. Comparison of Throughput on Llama2-70B offline inference

LLMBoost with AMD Instinct MI300X achieved a leap to 93,039 TPS in the server scenario and 103,182 TPS in the offline scenario—surpassing the performance of a 32x H100 setup. With each MI300X GPU estimated price of around $15,000–$17,000, compared to $32,000–$40,000 per H100— LLMBoost + AMD Instinct MI300X offers up to 62% savings in GPU cost while delivering superior inference throughput.[2]

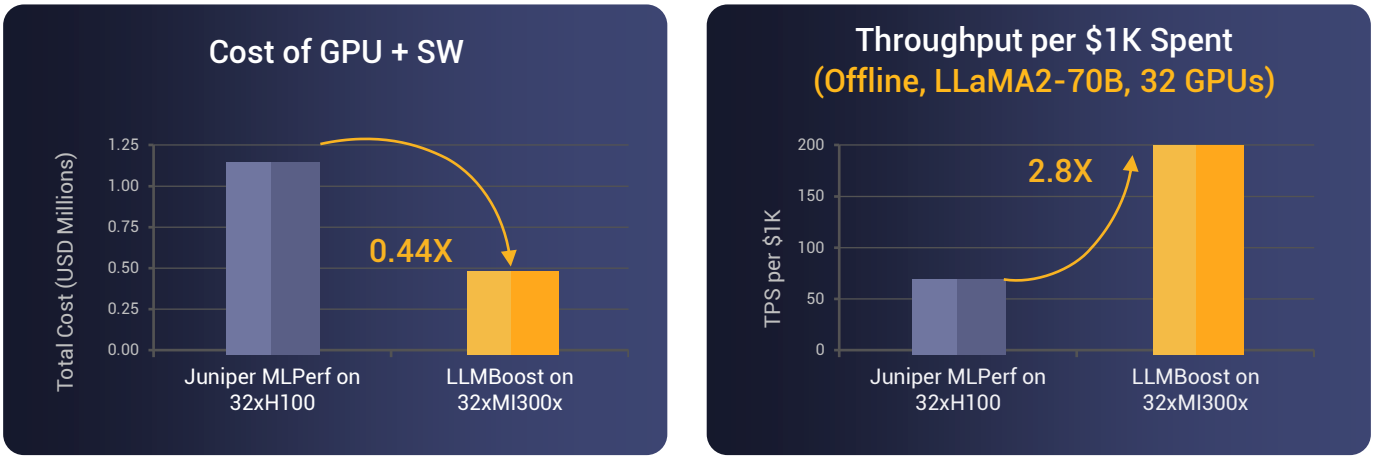

FIGURE 2. Estimated Cost vs. Performance Efficiency of 32-GPU Inference Systems

In terms of cost-efficiency, the LLMBoost with MI300X system delivers approximately 2.8× more inference throughput per $1,000 spent than the H100-based system, making it the clear choice for high-performance, budget-conscious deployments.

Highest Llama2-70B Offline Inference Results in MLPerf v5.0

Impressively, MangoBoost achieved the highest LLaMA2-70B offline inference performance in MLPerf v5.0 compared with all other submitters (e.g., Nvidia, Intel, Google). Compared to the previous 8-GPU MI300X submission[1], MangoBoost's latest MLPerf results showcase a linear performance scaling from 8 to 32 GPUs when serving Llama2-70B. These results demonstrate that LLMBoost can scale efficiently from a single-GPU to a distributed multi-node, multi-GPU production cluster.

Whether you're running a handful of inference threads or deploying a high-throughput LLM service, LLMBoost ensures you get the maximum efficiency at any scale.

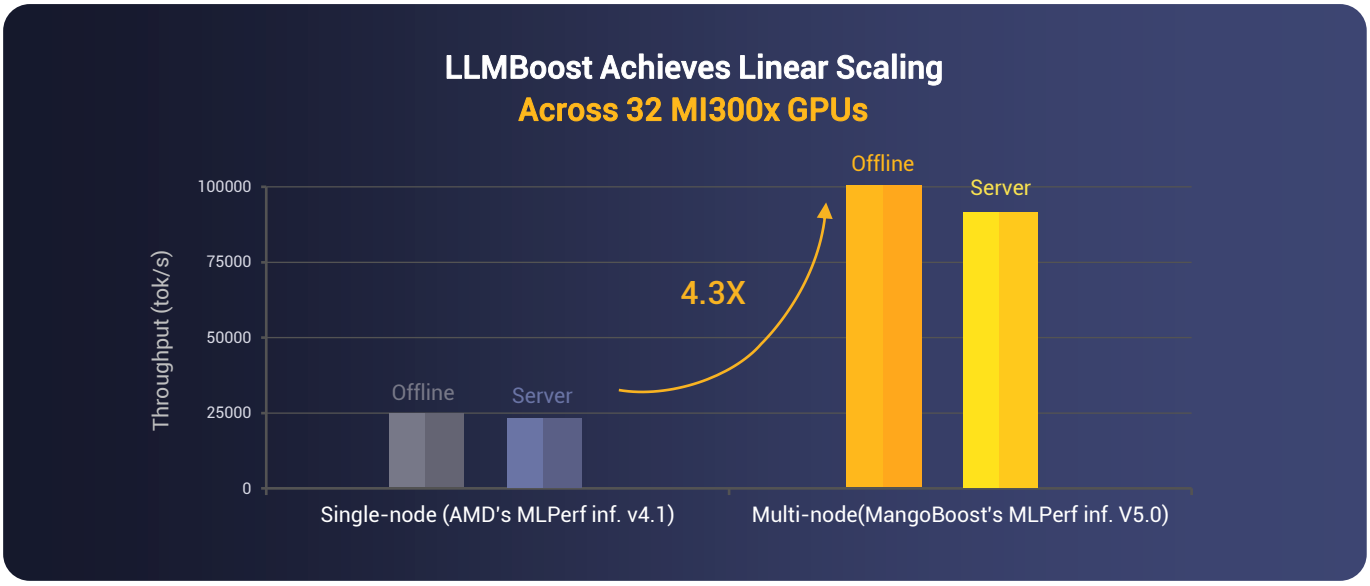

FIGURE 3. LLMBoost Achieves Linear Scaling Across 32 MI300x GPUs

Want to replicate our MLPerf v5.0 results? **Try it yourself on Azure**[3]
or follow the step-by-step instructions in our **MLPerf GitHub README**[4]

Additional Result Highlights on 8x Nvidia A100 GPUs on AWS

Beyond the specific MLPerf results mentioned above, LLMBoost has been extensively evaluated with many platforms and GPUs. As an example, in our evaluation using an 8×NVIDIA A100 GPU server from AWS, LLMBoost delivers up to 138× faster inference compared to Ollama, and outperforms both HuggingFace TGI and vLLM across all model sizes tested, including LLaMA3.1-70B, DeepSeek-R1-Distill-Qwen-32B, and LLaMA3.1-8B. In terms of cost-efficiency, LLMBoost also leads the pack with the lowest GPU cost per million tokens, reducing inference cost by over 99% compared to Ollama, and by over 30% even compared to vLLM on high-throughput workloads.

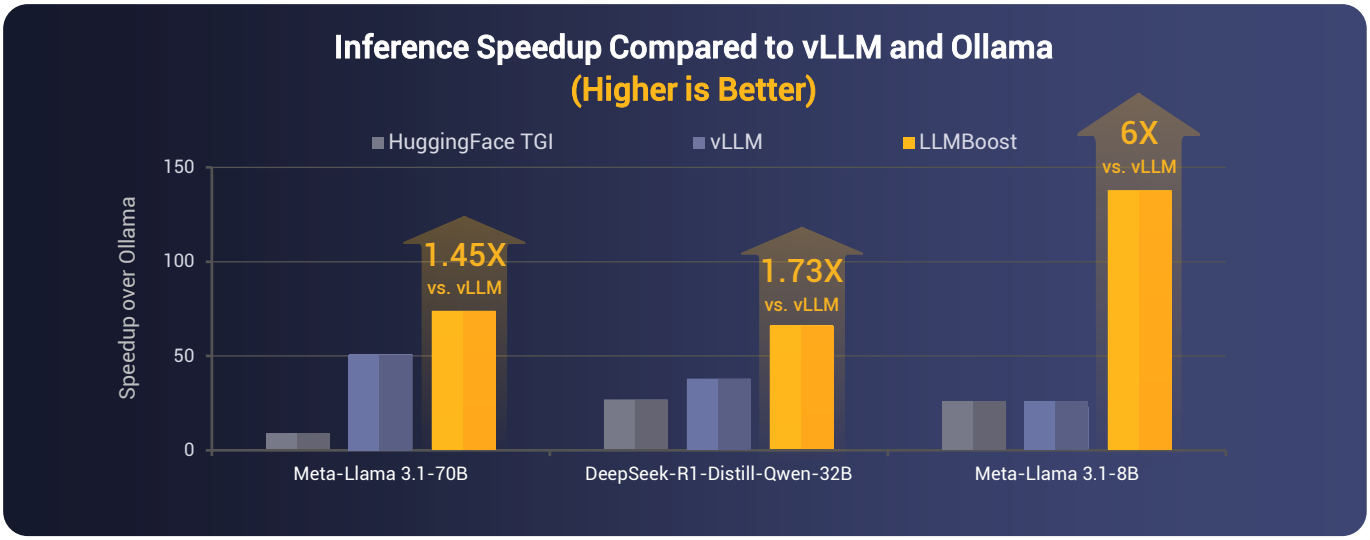

FIGURE 4. Inference Speedup Compared to vLLM and Ollama

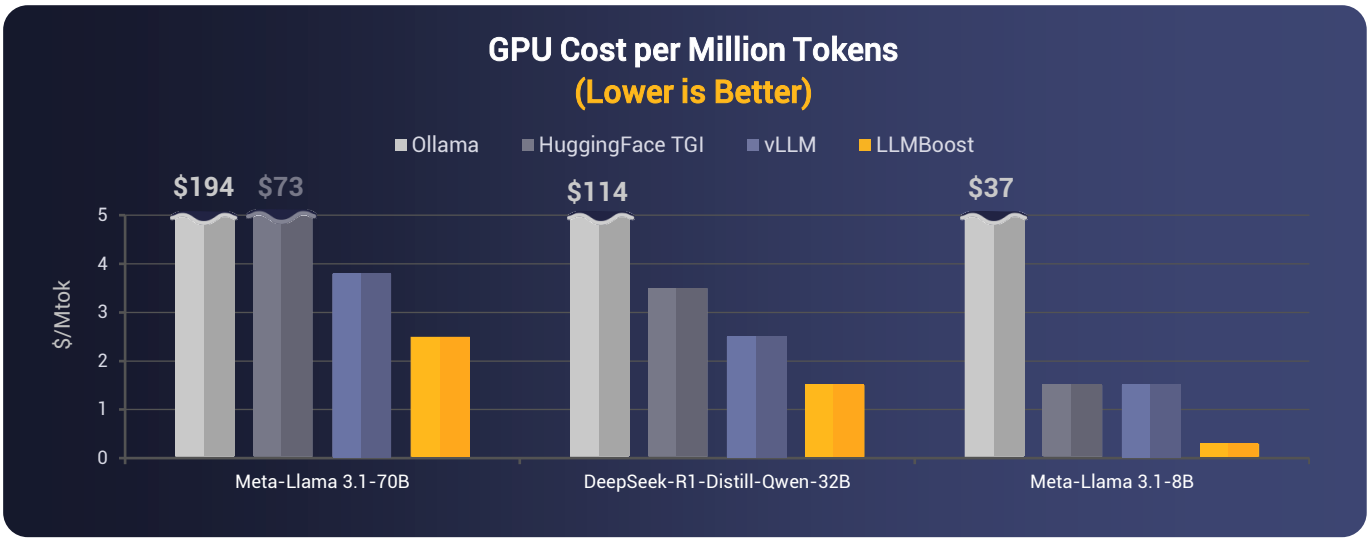

FIGURE 5. GPU Cost per Million Tokens

**Based on AWS pricing for 8×A100 GPUs at $33/hour (as of March 25, 2025).*

Whether you're optimizing for performance, cost, or deployment simplicity, LLMBoost offers unmatched efficiency and scalability across a wide range of LLMs.

From Research to Production in Minutes

LLMBoost bridges the gap between experimentation and deployment, allowing developers and data teams to serve powerful open-source models in minutes, not days. Whether you're testing a quantized Llama2-7B model, experimenting with multimodal models like Llava, or deploying a production-scale Llama2-70B cluster, LLMBoost delivers a consistent, streamlined experience at every scale.

Getting started is as simple as a running single command:

```
$ llmboost serve –model_name google/gemma-2-2b-it
[INFO] Initializing LLMBoost server...
```

MangoBoost AI Infrastructure Hardware Solutions based on DPUs

Beyond LLMBoost software, MangoBoost offers hardware acceleration solutions based on Data Processing Unit (DPUs) for AI and cloud infrastructure, such as:

- **Mango GPUBoost™** – RDMA acceleration for multi-node inference/training via RoCEv2
- **Mango NetworkBoost™** – Offloads TCP/IP stack to free up CPU resources
- **Mango StorageBoost™** – High-performance NVMe initiator/target for scalable AI storage

Try Mango LLMBoost™ on Cloud!

A ready-to-deploy, full-stack AI inference server offering unprecedented performance and flexibility

aws | Google Cloud | Azure

Try it now. If interested in our LLMBoost Solutions, please contact us at contact@mangoboost.io

References

[1] Benchmark Suite Results for MLPerf Inference: Datacenter

- https://mlcommons.org/benchmarks/inference-datacenter/

[2] Tom's Hardware – H100 vs. MI300X Pricing

- https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000

[3] Mango LLMBoost for AMD MI300x on Azure Marketplace

- https://azuremarketplace.microsoft.com/en-us/marketplace/apps/mango_solution.mango_llmboost_amd?tab=Overview

[4] MangoBoost's MLPerf GitHub README

- https://github.com/mlcommons/inference_results_v5.0/blob/main/closed/MangoBoost/README.md