WEKA

Top 7 Reasons You Shouldn't Wait on Inference

And Why WARRP Gets You There Faster

In today's AI-driven world, inference is where real value is created—powering everything from chatbots and copilots to recommendation engines and autonomous agents. But building a scalable, high-performance inference stack isn't easy.

That's why we built the WEKA AI RAG Reference Platform—or WARRP.

WARRP is a modular, production-grade architecture that radically simplifies the deployment of Retrieval-Augmented Generation (RAG) pipelines. It brings together high-speed storage, pre-integrated models, GPU orchestration, and vector database tools into a single, performance-optimized stack—deployable in just one line.

Here's why WARRP is the fastest path to scalable inference:

1 Slow Inference Hurts Engagement and ROI

Latency kills. If your model response times drag, so does user satisfaction—and GPU efficiency.

2 RAG Grounds LLMs

Retrieval-Augmented Generation adds referenceable citable knowledge to your models—but only if your data stack can keep up.

3 DIY Inference Pipelines Are a Headache

Orchestration, storage, models, vector DBs... Building your own stack is hard. WARRP gives you an integrated architecture that just works.

4 Built on NeuralMesh for Blazing Performance

WARRP is built on NeuralMesh by WEKA—so you get ultra-low-latency, high-throughput data access and unmatched parallelism.

5 Optimized for NVIDIA Ecosystems

WARRP supports NVIDIA NIMs, NV Ingest, Triton, and popular models like DeepSeek and Llama v3—right out of the box.

6 Cloud, On-Prem, or Both—Same Stack

WARRP runs identically across cloud and on-prem environments. That means no trade-offs in performance or portability.

7 Production-Ready from the Start

Forget proof-of-concepts that never scale. WARRP is built for real-world RAG, inference, and agentic workloads from day one.

Want to see how WARRP accelerates your inference workflows? Let's talk.

WEKA