DOI: doi.org/10.47363/JAICC/ICMLAIDS2026/2026(5)18

International Conference on Machine Learning, Artificial Intelligence and Data Science (ICMLAIDS 2026)

Conference Proceedings March 20 - 21, 2026 - Virtual

The AI Performance Illusion: A Verification Perspective on GPU, Cloud, and Deployment Variability

Santosh Appachu

Independent Researcher, Texas, USA

Abstract

Enterprises often assume that AI performance is primarily determined by the model and the GPU. In reality, production deployments repeatedly reveal an "AI performance illusion": identical models can behave very differently across GPU types, cloud platforms, and deployment environments, even when the hardware appears similar. These performance gaps are frequently misunderstood as model issues, while the root causes originate from lower-level execution and infrastructure layers such as memory movement, kernel scheduling, runtime configuration, data pipelines, container environments, and orchestration behavior.

This talk introduces a verification-driven perspective for diagnosing and preventing performance variability in modern AI systems. Instead of relying on trial-and-error tuning, we treat AI performance as a system property that must be validated using repeatable measurement methodology, controlled experiments, and infrastructure-aware observability. The session breaks down real-world causes of GPU underutilization, throughput collapse, and latency spikes, covering CPU - GPU interaction bottlenecks, dynamic batching behavior, precision modes, multi-process interference, software stack inconsistencies, and cloud deployment drift.

Attendees will leave with a practical blueprint for verifying AI performance across environments, ensuring reproducibility, and building deployment pipelines that deliver predictable throughput, stable latency curves, and cost-efficient scaling.