

AWARDED TO

Tomas Ibarra

NVIDIA Triton Inference Performance Tuning

Profiles NVIDIA Triton workloads, identifies latency and throughput constraints, and tunes batching, concurrency, and model-serving resources against measurable targets.

GRADE	SCORE	ASSESSED
Expert	99 / 100	9 September 2026

Rafael Duarte
Lead Technical Assessor, GigsHarbor

GH-2026-ED5ADE
Issued by GigsHarbor on the strength of documents
the holder provided and an assessment by our expert.