Production LLMOps on Azure: NHL goes from notebooks to a service.

Microsoft Azure AI Solutions Engineer · In-flight · Telco, Media & Gaming portfolio

CUSTOMER	COMPUTE	SERVING	STATUS
NHL major sports league	**A100 multi-GPU** PyTorch DDP + FSDP	**vLLM** production latency budget	**In flight** customer-operated

THE SETUP

A major professional sports league wanted to move from GenAI experimentation to something that could run in production at the scale their content and fan-facing workflows demanded. Models worked in notebooks. Proofs-of-concept demoed well. They didn't have distributed training that stayed up, inference serving that hit latency targets under real load, a cost story that worked when the bill came in, or a dedicated MLOps team to build any of it. The gap between "impressive demo" and "reliable service" is where most GenAI initiatives stall.

THE ARCHITECTURE

`Training` **Distributed multi-GPU on A100 clusters.** PyTorch DDP and FSDP for foundation-model fine-tunes. Orchestration hardened against the failure modes that surface at scale: stragglers, checkpoint corruption, cluster-level restarts.

`Serving` **vLLM with Azure-native autoscaling.** Tuned for throughput-per-dollar on the latency budget the product required. Cost-per-inference is a first-class metric, not an afterthought.

`Eval` **Regression harness, fail-closed guardrails.** Prompt and config regression tests, quality metrics tracked over time, guardrails that fail closed rather than open when conditions degrade.

`Cost obs` **Per-request attribution + deploy gate.** Per-request cost attribution, drift alerting, and a deploy gate so the team sees the cost impact of a release before it ships.

`Handoff` **Customer team operates the stack.** Paired directly with the customer's platform engineers on the rollout. The pattern can be operated by the team after handoff — not dependent on me long-term.

OUTCOME

COST-PER-INFERENCE	PATTERN	OPERATING MODEL
`Materially down` `consistent with portfolio range`	**`Reusable`** `across Telco, Media, Gaming`	**`Customer-led`** `post-handoff`

> **What this proves.** Production LLMOps is a different discipline from notebook ML. Distributed training that survives a week, inference that holds latency under real traffic, and a bill the CFO can read — that's the bar.