

Fine-Tuning Small Language Models for the Edge

Specialized intelligence for scale on Dell PowerEdge servers and Broadcom networking

Edge AI delivers useful answers where work happens: a technician servicing remote equipment, an operator responding to a plant alarm, a pilot receiving an onboard advisory. **Specialized small language models** meet that constraint: fine-tune a compact model to the domain, develop and validate it centrally, then reuse it across many edge deployments.

KEY FINDINGS

+5.2 / +6.0 pts

DOMAIN SPECIALIZATION

Accuracy gains for two fine-tuned 14B models on a 906-question aviation maintenance assessment, in four-bit deployment format.

96% / 94%

PRODUCTIVE SCALING

Concurrency efficiency running independent fine-tuning jobs across 8 GPUs (PowerEdge R770) and 16 GPUs (PowerEdge XE7745).

1.53x / 1.58x

CLUSTER SCALING

Distributed-training throughput gain from adding a second server over Broadcom 400GbE, on R770 and XE7745 respectively.

WHAT THE CASE STUDY ESTABLISHES

Anomaly detection and a rules engine establish the event; the model explains it and answers follow-up questions using retrieved maintenance information and aircraft context.

MODEL, Q4_K_M	BASE	FINE-TUNED	CHANGE
QWEN3 14B	71.5%	76.7%	+5.2 PTS
PHI-4 14B	71.4%	77.4%	+6.0 PTS

Pre-sweep test accuracy compared with post-training results. Other experiments regressed; the results support targeted specialization, not a guarantee.

27 of 30

KNOWLEDGE ≠ BEHAVIOR

Advisory prompt cases passed by the Qwen3 fine-tune; other candidates passed 30 of 30. Domain and application tests are separate gates.

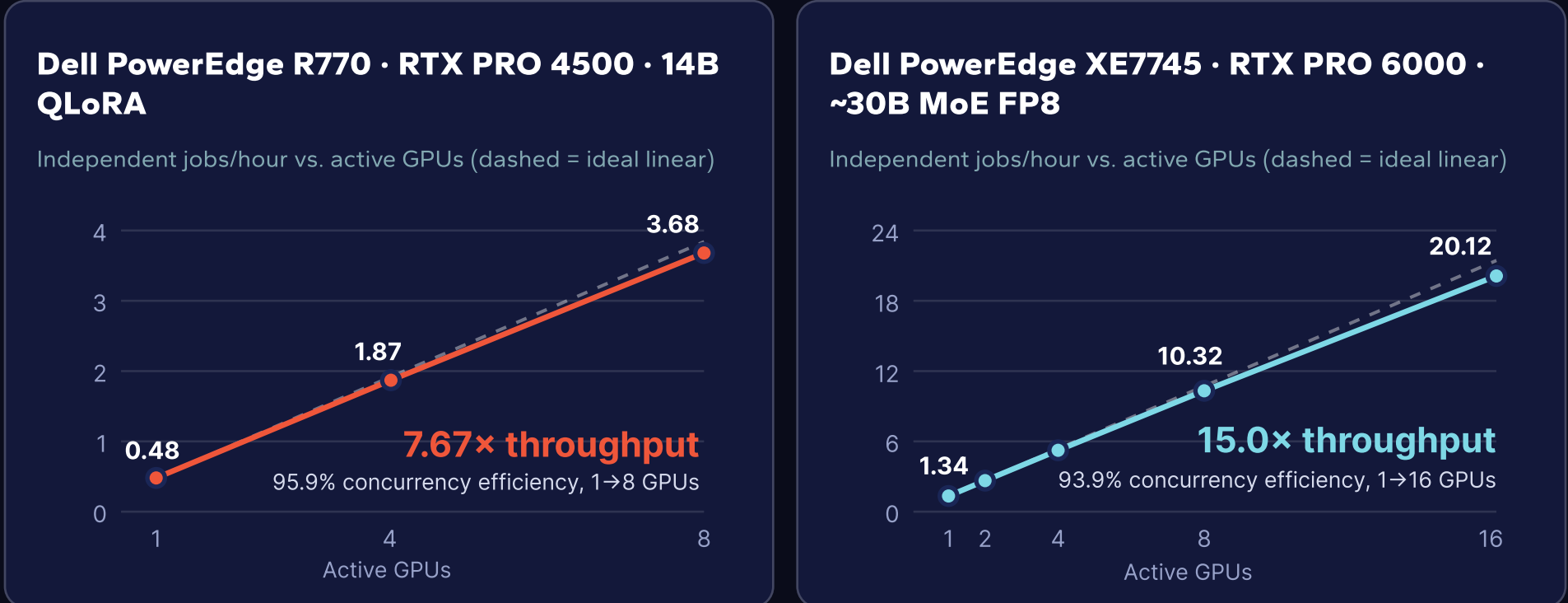
-8.3 pts

COMPRESSION CAN COST ACCURACY

Gemma 4 26B-A4B fine-tune: 79.6% at Q4_K_M, 71.3% at Q3_K_S. Qualify the exact deployed artifact.

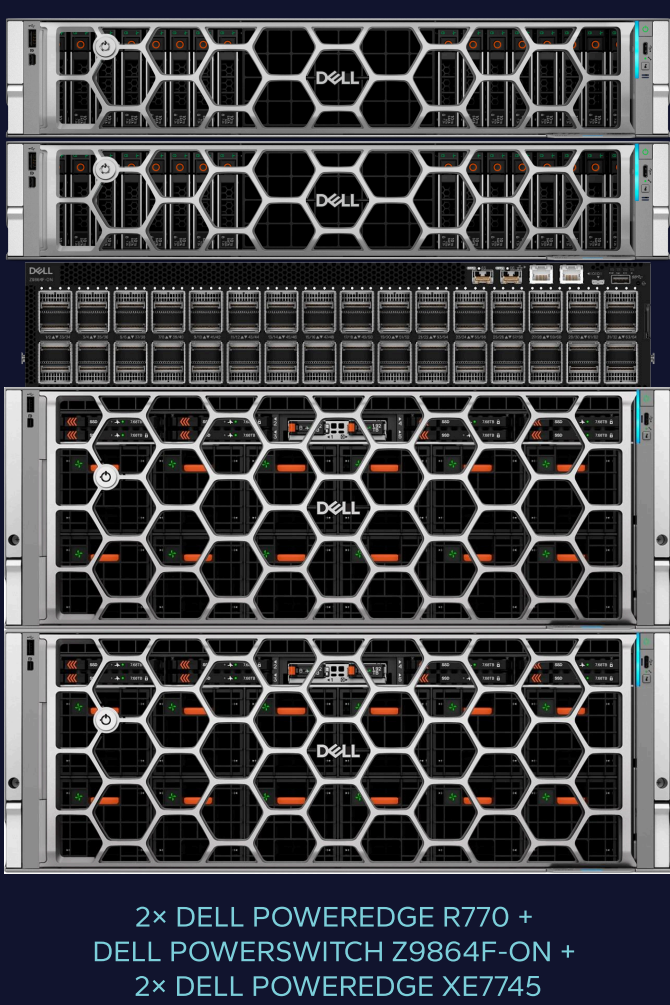
TWO MEASURES OF SUCCESS

Independent-job concurrency (one fine-tuning job per GPU) expands how many candidates, domain variants, and repeat validations complete per day. **Distributed training** (several GPUs cooperating on one job) shortens time to target quality. Both were measured on Dell PowerEdge R770 and XE7745 servers.



Independent jobs scale with modest per-job slowdown: individual job duration rises ~4.3% on R770 and ~6.5% on XE7745 between the endpoints. Workloads differ, so job rates are not a direct platform comparison.

THE TESTED MODEL-DEVELOPMENT STACK



2× DELL POWEREDGE R770 +
DELL POWERSWITCH Z9864F-ON +
2× DELL POWEREDGE XE7745

- Dell PowerEdge R770**

 - Up to 4×NVIDIA RTX PRO 4500 Blackwell Server Edition, 32GB each
 - Broadcom BCM57608 400GbE RoCEv2 network controllers, PCIe 5.0
 - NVIDIA driver 580.95.05 · Linux kernel 6.8.0-87 · Accelerate, xFormers
 - Workload: independent 14B QLoRA jobs
- Dell PowerSwitch Z9864F-ON — The Fabric**

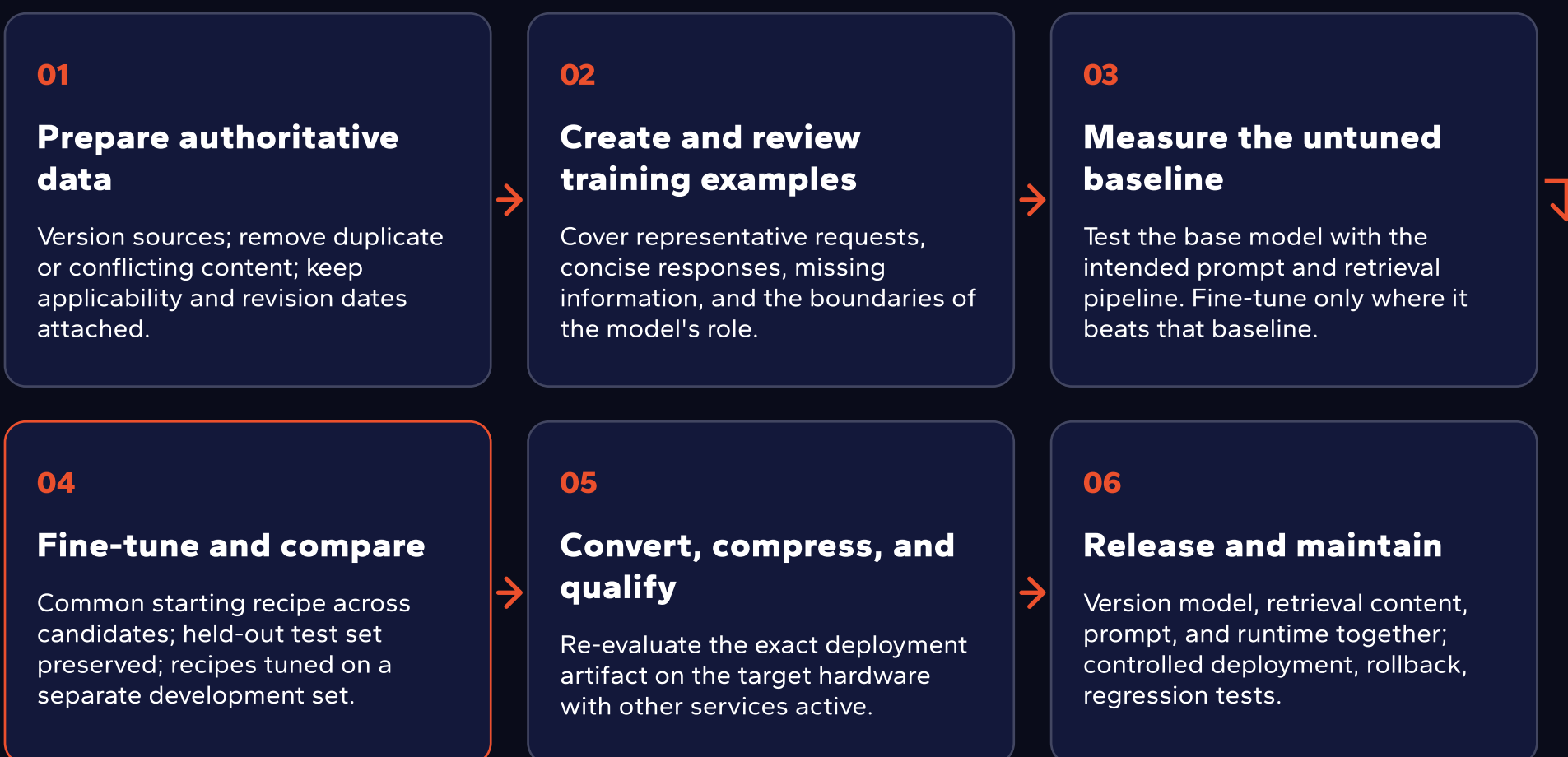
Built on Broadcom Tomahawk 5 silicon, the switch interconnects the servers over 400GbE with RoCEv2. Each server's BCM57608 adapters are configured as 1×400GbE links, carrying shared datasets, base models, and checkpoints for independent jobs and gradient synchronization for distributed training.
- Dell PowerEdge XE7745 — The Compute**

Each XE7745 node in the tested cluster:

 - Up to 8× NVIDIA RTX PRO 6000 Blackwell Server Edition, 96 GB each (768 GB per node)
 - Broadcom BCM57608 400GbE RoCEv2 network controllers, PCIe 5.0
 - NVIDIA driver 580.95.05 · Linux kernel 6.8.0-87 · unsloth, TRL, PEFT, Accelerate, bitsandbytes, xFormers

A REPEATABLE PATH FROM DOMAIN DATA TO EDGE RELEASE

Start with the **deployment contract** — task, authoritative sources, permitted outputs, acceptable failure behavior, latency target, and device resource ceiling — before choosing a model. Evaluation then becomes an acceptance decision, not a search for the highest generic score.



Reviewed feedback and updates return to data preparation — fine-tune when an interaction pattern changes; refresh retrieval when a document changes

Deploy the smallest model that reliably meets the operating requirement

Develop specialization centrally, where teams share infrastructure, compare candidates, and repeat evaluation; qualify the complete application on its target device; then spread the cost of specialization across the fleet. Dell PowerEdge R770 and XE7745 servers with Broadcom networking supply the development capacity that keeps an edge model qualified as requirements change.