



EXECUTIVE SUMMARY

Fine-Tuning Small Language Models for the Edge

Building Specialized
Intelligence for Scale

AUTHOR

Brian Martin

AI Data Center Performance | Signal65

IN PARTNERSHIP WITH

DELLTechnologies

OCTOBER 2026

For enterprises deploying AI into aircraft, industrial sites, or field-service equipment, model size has a direct operating cost. More memory, longer responses, and dependence on a network connection can increase device requirements or interrupt useful work. Across a fleet, those constraints repeat at every deployment.

Specialized small language models offer a practical way to bring domain knowledge to these environments. Fine-tuning adapts an existing model to the vocabulary and behavior a task requires. Local retrieval supplies current evidence, while application controls keep the model within its intended role.

The opportunity is to develop specialization centrally and reuse it wherever work happens. 550 Aero's aviation experiments demonstrate measurable gains in compact-model accuracy. Dell PowerEdge systems with Broadcom networking provide the capacity to compare candidates, build domain variants, and refresh models as requirements change.

Key Highlights



Domain specialization: 5.2–6.0 percentage point gains for two compact models



Productive scaling: 96% (R770, 8 GPUs) and 94% (XE7745, 16 GPUs) concurrency efficiency



Cluster scaling: distributed training delivered **1.53x throughput on R770** and **1.58x on XE7745**

The Economics of Specialization

A field technician needs an approved procedure for a specific machine. An operator needs an explanation of an alarm in the context of local operating limits. These bounded tasks create room for specialization. Lower memory requirements can make smaller devices practical or leave capacity for other services, while local execution maintains availability during network loss.

The business case spans development and operation. Central infrastructure can spread the cost of adapting and validating a model across many deployments. At the edge, savings in memory, energy, and response time can accumulate across devices and repeated interactions. The relevant measure is cost per accepted task, including the cost of unsuccessful experiments, failed responses, and retries.

That advantage depends on qualification. A model that scores well on domain knowledge must also follow application rules and meet memory and latency targets on the intended device. The investment decision is to select the smallest model that reliably completes the work, supported by the development capacity to keep it qualified.

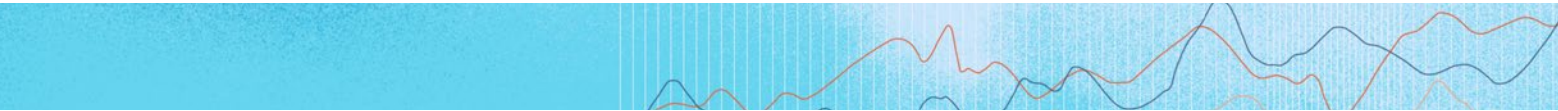
Measured Gains in Domain Capability

550 Aero evaluated two 14B models on a 906-question aviation maintenance assessment. Fine-tuning raised accuracy by 5.2 and 6.0 percentage points in a four-bit deployment format, showing that focused adaptation can add useful domain capability to compact models.

Model and evaluation format	Base accuracy	Fine-tuned accuracy	Change
Qwen3 14B, Q4_K_M	71.5%	76.7%	+5.2 percentage points
Phi-4 14B, Q4_K_M	71.4%	77.4%	+6.0 percentage points

Pre-sweep test accuracy compared with post training results

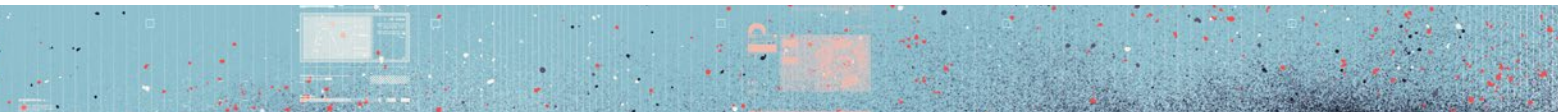
The results also define the limits of the claim. Other experiments regressed, and strong domain scores did not always yield reliable application behavior. The reported Qwen3 fine-tune passed 27 of 30 advisory prompt cases, while other candidates passed 30 of 30.



Development Capacity That Scales

The infrastructure results show how organizations can expand specialization by running more independent jobs. Across two Dell PowerEdge R770 servers, eight RTX PRO 4500 Blackwell Server Edition GPUs delivered 3.68 jobs per hour, or 7.67x the single-GPU baseline. Across two PowerEdge XE7745 servers, 16 RTX PRO 6000 Blackwell Server Edition GPUs delivered 20.12 jobs per hour, or 15.0x the baseline. These gains increase capacity to evaluate candidates and maintain variants. The workloads differ, so their absolute job rates are not a direct platform comparison.

Broadcom BCM57608 adapters configured as 1x400GbE connect the servers through Dell PowerSwitch Z9864F-ON switches powered by Broadcom Tomahawk 5. Adding a second server increased distributed-training throughput by 1.53x on R770 and 1.58x on XE7745 relative to their measured single-server baselines. Planning compute and networking together helps translate GPU investment into productive development capacity.



Conclusion

The enterprise opportunity is to extend centrally developed expertise across a fleet of constrained devices. Dell compute and Broadcom networking support the development work; qualified compact models deliver useful answers where that work is needed. Evaluate the return through cost per accepted task, energy per accepted edge task, and time from approved change to qualified release. Together, these measures connect infrastructure spending to the reliability and efficiency of the deployed application.

Important Information About this Report

CONTRIBUTORS

Brian Martin

AI Data Center Performance | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com or contact research@signal65.com



IN PARTNERSHIP WITH



Read the full report: [Fine-tuning Small Language Models for the Edge](#)



CONTACT INFORMATION

Signal65 | signal65.com