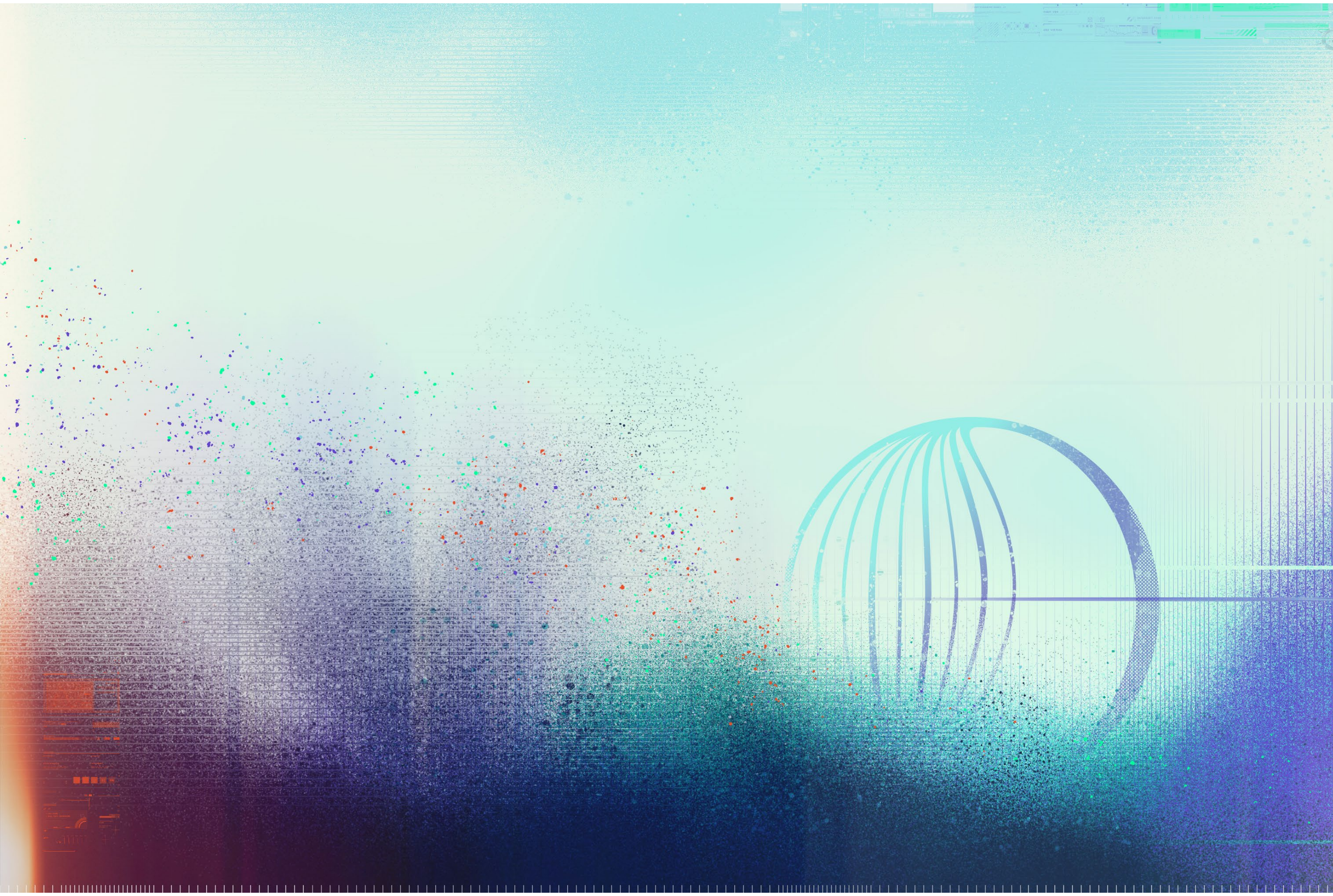
signal65



EXECUTIVE SUMMARY

Evaluating Agentic AI

Dell AI Solutions with Open Models

AUTHORS

Mitch Lewis
Performance Analyst | Signal65

Brian Martin
AI Data Center Performance | Signal65

DECEMBER 2025

IN PARTNERSHIP WITH

KAMI WAZA

DELL Technologies

Overview

AI has rapidly shifted from experimental concept to mission-critical capability for modern enterprises. As organizations move beyond single-turn question answering and toward **agentic AI** to autonomously complete multi-step business tasks, new evaluation methods are required to determine how well models can actually do work, not just reason in isolation.

Traditional benchmarks fail to capture this. Many are static, prone to data leakage, and optimized for reasoning over execution. They do not reflect real enterprise workflows, where models must repeatedly call tools, extract data, handle files, navigate directory structures, query databases, and return validated output.

To address this gap, Signal65, in partnership with Kamiwaza, created a new benchmarking methodology: the Kamiwaza Agentic Merit Index (KAMI). This dynamic benchmark evaluates practical agentic task performance across LLMs running on modern on-premises infrastructure including Dell PowerEdge XE7745 servers with Broadcom 400 GbE RoCEv2 fabrics, reflecting how enterprises deploy AI for secure, cost-efficient, latency-sensitive workloads.

Key Highlights

31 Models tested with **over 170,000 conversations**

5.5 billion tokens processed to date in testing

Qwen3-235B-A22B Instruct-2507-FP8 led with an **88.8% mean accuracy score**

The KAMI Benchmark

KAMI creates a uniquely randomized test environment for every run

Elements:

- **Sandbox file and database environments** randomized per sample
- **Dynamic entity substitution,** generating unique instructions
- **Runtime-generated ground-truth answer keys**
- **Tool-calling loops through an agentic server**
- **Multi-step enterprise tasks,** not single-shot answers

Examples:

- Creating files and directory structures
- Extracting exact values from CSV and text files
- Running SQL queries against randomized business databases
- Formatting results in strict JSON or text schemas

This approach produces contamination-free, execution-focused evaluation of genuine agentic capability.

Key Findings

1. Model Size Strongly Correlates with Agentic Accuracy

Across the 31 models evaluated, very large models (>100B) performed best overall. Large models (60-75B) excelled at filesystem and text extraction operations. Medium models (10-50B) were surprisingly competitive, and small models (<10B) performed poorly across nearly all test categories.

2. Quantization (FP8) Had Minimal or No Negative Impact

In several FP8 vs full-precision pairs, the **quantized versions slightly outperformed** their full-weight counterparts—a strong signal that efficiency-oriented FP8 deployment does not degrade agentic task performance.

3. Thinking Models Outperformed Non-Thinking Counterparts

Across Qwen3 models, "thinking mode" consistently improved accuracy, especially in database, CSV, and multi-step retrieval, as well as complex filesystem and instruction-following workflows. Surprisingly, providing hints significantly narrowed the gap, highlighting a potential hybrid strategy for organizations balancing cost with accuracy.

The full results of the KAMI v0.1 benchmark can be seen in Figure 1 below.

Rank	Model	Mean Accuracy Score
1	Qwen3-235B-A22B-Instruct-2507-FP8	88.8 %
2	Qwen3-235B-A22B-Instruct-2507	88.4 %
3	Claude-3.5-Haiku-20241022	75.9 %
4	Llama-4-Maverick-17B-128E-Instruct	74.6 %
5	Llama-3.3-70B-Instruct-FP8-KV	74.5 %
6	Llama-3.1-70B-Instruct	73.4 %
7	Llama-4-Maverick-17B-128E-Instruct-FP8	73.1 %
8	Qwen3-30B-A3B (thinking mode)	72.7 %
9	Llama-3.3-70B-Instruct	71.6 %
10	Qwen2.5-72B-Instruct	71.1 %

Figure 1: *KAMI v0.1 Benchmark Top 10 Results*

The Dell PowerEdge XE7745 delivers enterprise-grade AI infrastructure, providing a high-memory, multi-GPU platform that makes Agentic AI practical at scale. Equipped with dual AMD EPYC processors, up to 3 TB of DDR5 memory, support for eight NVIDIA L40S, H100, H200, or RTX Pro 6000 GPUs, and eight Broadcom BCM57608 400 GbE network controllers, each node delivers exceptional bandwidth and parallelism for large-context inference and high-volume retrieval. Clustered with Dell Z9864F-ON switches running Dell Enterprise SONiC within a RoCEv2 fabric, these solutions deliver deterministic, low-latency, and low-jitter performance across diverse workloads. This balance of memory capacity, compute density, and network efficiency enables businesses to deploy Agentic AI systems for the most demanding reasoning workloads.

Conclusion

Understanding the agentic capability of models is important for enterprise organizations considering multi-agent applications. Tools like the KAMI benchmark can provide context and data to help understand the strengths and weaknesses of various LLMs in real agentic scenarios.

Agentic AI will define new enterprise AI applications. Organizations that invest in the right infrastructure today, combining the computational power of the Dell PowerEdge XE7745, the networking efficiency of Broadcom BCM57608 400 GbE NICs, and the sophistication of evolving models and agentic designs, will be best positioned to harness the full potential of AI-driven knowledge systems. This positioning will enable them to drive innovation and maintain competitive advantage in an increasingly AI-powered business landscape.

Download the full report here.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65
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

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 an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Acknowledgments

The authors would like to thank Kamiwaza for the KAMI benchmark and Dell Technologies for providing access to the PowerEdge XE7745 hardware platform and technical expertise. Special recognition goes to the engineering teams at Broadcom for their collaboration on network optimization strategies, and to the open-source community behind vLLM for their contributions to this field.

CONTACT INFORMATION

Signal65 I signal65.com