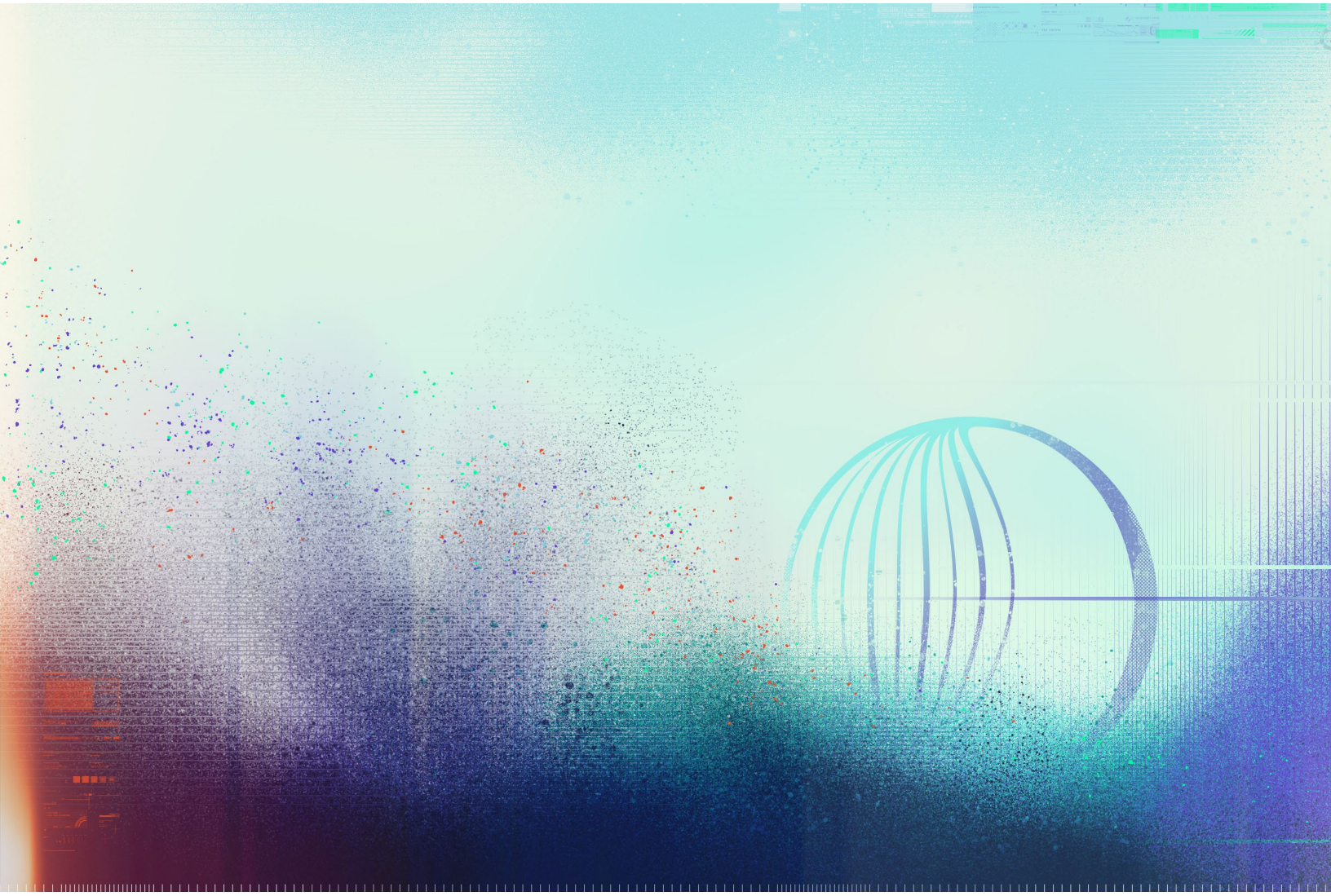
signal65



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

KAMIWAZA

DELL Technologies

Executive Summary

AI has rapidly moved from a emerging technology to a top priority for enterprise organizations. Continual advances in model development produce larger, more capable, and more reliable LLMs. For enterprise adoption, agentic applications provide tangible business value, enabling AI to solve challenges and complete meaningful business tasks with minimal human involvement.

As agentic AI becomes the center of enterprise strategy, organizations increasingly recognize the need to run AI on premises to maintain data sovereignty, preserve operational control, and achieve predictable, cost-effective performance. Yet, evaluating LLM usefulness and accuracy on agentic tasks remains challenging, as most existing benchmarks focus on reasoning abstractions rather than a model's ability to execute real enterprise workflows. Static benchmarks can also leak into training data, turning them into exercises of memorization rather than true capability tests.

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**

To address this gap, Signal65 and Kamiwaza established a new AI benchmark to measure model performance on enterprise-focused agentic tasks, validated on modern on-premises infrastructure including Dell AI servers and Broadcom high-performance Ethernet fabrics. These open, standards-based environments reflect how organizations deploy AI today: securely under their control and optimized for both performance and cost. Testing covered 31 models and over 170,000 test conversations, resulting in over 5.5 billion tokens processed.

Key findings:

- **Model Size:** In general, accuracy was seen to improve with model size, with the highest scores attributed to very large models with over 100B parameters. Small models (<10B parameters) showed a clear deficiency across most agentic tasks. Some models in the 30 to 100B parameter range, however, such as Llama-3.1-70B-Instruct and Qwen3-30B-A3B (thinking mode) outperformed much larger models, demonstrating compelling options for organizations with limited infrastructure.
- **Quantization:** FP8 quantization does not appear to have adverse effects on agentic capabilities. Across FP8 quantized and full weight model pairs tested, the FP8 variations consistently achieved similar or even slightly greater accuracy.
- **Thinking:** Models with thinking capabilities were generally found to be more accurate in achieving agentic tasks than similar non-thinking models. Non-thinking models, however, became highly competitive when provided basic hints and context clues, offering a possible alternative to the high token usage and cost associated with thinking models.

Challenges with Agentic Benchmarking

Agentic AI has quickly become the focus of enterprise AI adoption. By leveraging agents, AI can perform useful enterprise tasks, from simple routine tasks to complex multi-step operations. As enterprises begin building such agentic systems, evaluation of LLMs becomes a crucial component. The chosen LLM will have a direct impact on the accuracy, effectiveness, and efficiency of the agent. While there are several benchmarks currently available to evaluate AI models, existing approaches use flawed methods for evaluating models for true agentic use cases.

Most existing AI benchmarks are static, leading to data contamination and memorization issues. With static benchmarks, the benchmark itself can easily be introduced into a model's training data, intentionally or unintentionally. This invalidates the challenge of the benchmark, instead resulting in a test of memorization. When evaluating static benchmarks, it can't be known if a model is performing well due to its own merit or due to previous exposure to the benchmark. While the creation of new benchmarks temporarily solves this problem, it is not a scalable approach to accommodate ongoing model development.

Additionally, existing benchmarks do not accurately represent agentic use cases. Most benchmarks only evaluate single-shot question and answer responses. Agentic workflows, on the other hand, often involve multi-step inference and tool calling to complete specific enterprise related tasks, such as querying a database, evaluating the data, and formatting a result.

These limitations can result in misleading or uninformative results, making it challenging for enterprises to select appropriate models to support agentic applications, leading to inaccurate results and incomplete tasks. These critical mistakes can cause costly disruptions and potentially negate the advantages of leveraging agentic AI. For complex, multi-agent applications, the importance of model selection is essential, as inaccuracy in a single agent can compound throughout the application, impacting other agents and the final quality and reliability.

About the KAMI Benchmark

To address the growing need to evaluate the agentic capabilities of AI models, Signal65 have partnered with Kamiwaza to develop a new AI benchmark, known as the Kamiwaza Agentic Merit Index (KAMI).

KAMI differentiates itself from other AI benchmarks with a dynamic test suite that measures the completion of real agentic tasks. Unlike static AI benchmarks, KAMI randomizes each question and generates a unique ground truth answer key at runtime, preventing models from memorizing the tests during training. In addition to preventing memorization, KAMI also goes beyond measuring simple reasoning capabilities, with tests designed to evaluate a model's proficiency in completing actual enterprise tasks.

KAMI requires models to reason through enterprise-oriented tasks, such as answering business questions by extracting information from CSV files or databases. These tasks accurately represent common, real-world agentic workflows that involve loops of LLM inference and tool calling.

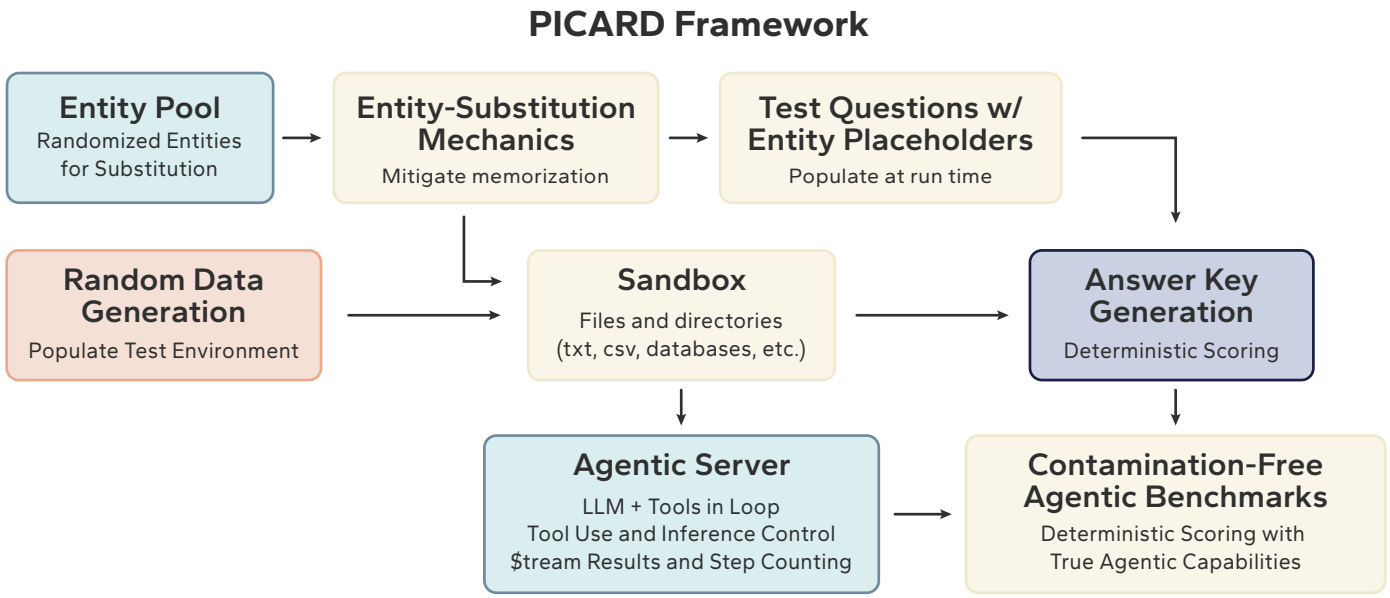


Figure 1: *PICARD Framework Architecture*

KAMI, based on the PICARD framework, creates a dynamic, randomized test suite targeted at specific enterprise workloads. Key components of KAMI include:

- **Sandbox Environment:** To enable real, agentic tasks, KAMI creates a sandbox environment for LLMs to do work as needed, including writing files, or connecting to databases.
- **Multi-layer Randomization:** KAMI creates unique, dynamic questions by deploying two layers of randomization.
  - **Entity Substitution:** randomizes the relevant entities in each benchmark question from a pool of possible entities, including file names, directories, and database tables.
  - **Data Generation:** randomizes the data, such as directories, files, and databases within the test environment.
- **Answer Key Generation:** All responses are graded against a ground truth answer key, generated at run time.
- **Agentic Server:** To evaluate real-world agentic tasks, an agentic server is deployed to enable LLMs to iteratively select tools, execute tools, and evaluate results in a loop for complex tasks.

An example of a question in the KAMI benchmark can be seen in Figure 2 below:

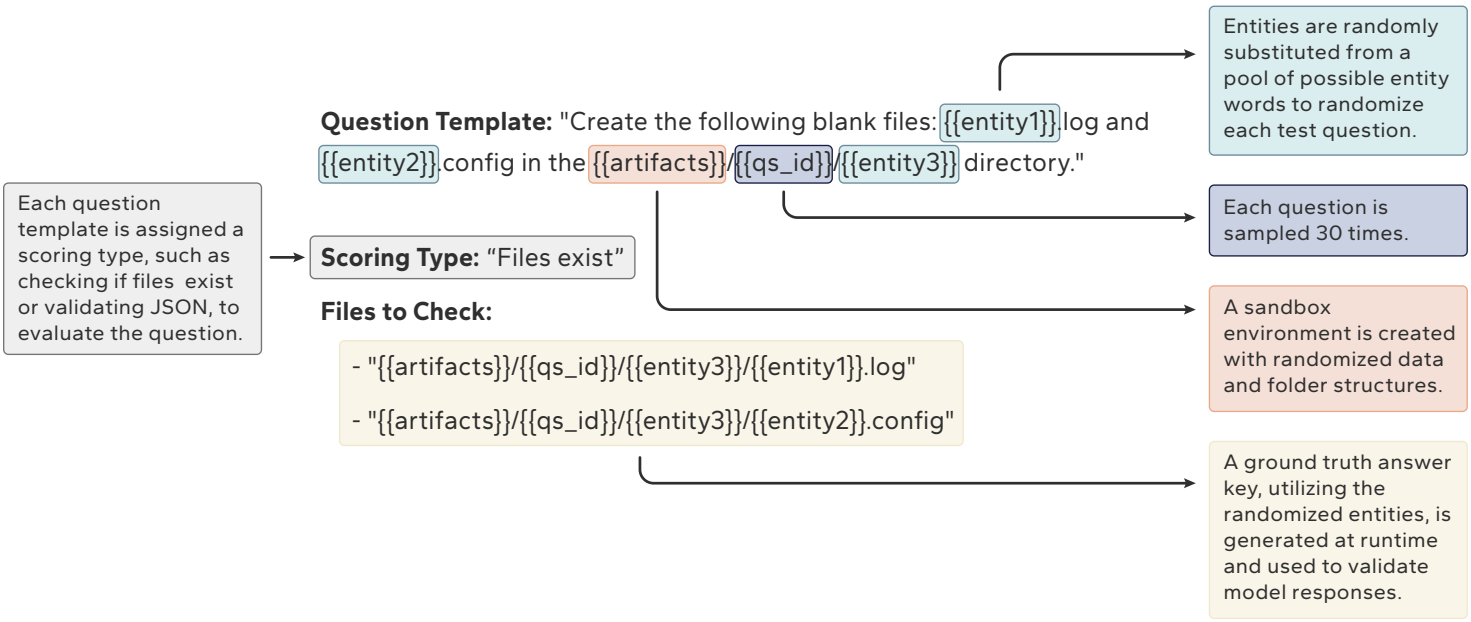


Figure 2: *Question Template Overview*

Figure 3 demonstrates this question populated with randomized data, along with a correct answer and a possible incorrect answer.

Example:

"Create the following blank files: crimson.log and whisper.config in the test_artifacts/q201_s20/ancient directory."

Correct Answer:

- "test_artifacts/q201_s20/ancient/crimson.log"
- "test_artifacts/q201_s20/ancient/whisper.config"

The correct answer created the properly named files in the correct directories.

Incorrect:

- "test_artifacts/q201_s20/ancient/crimson/ancient.log"
- "test_artifacts/q201_s20/ancient/whisper"

Any answer that doesn't match the expected answer key is incorrect. In this example, the first file is incorrectly named and placed in the wrong folder. The second file is in the correct directory and matching the entity name, but missing the proper file extension.

Figure 3: *Randomly Generated Question Example*

This example demonstrates a simple question within the Kami benchmark, which tests if an agent can create basic files and place them in the correct directory. The full benchmark contains many more complex questions involving gathering randomized information from databases, CSV files, and text files. An example of a database question can be seen in Figures 4 and 5. Across the various test cases, models were found to generate incorrect responses for wide range of reasons. Common challenges included incorrect or random tool usage, semantic confusion, and mishandling numerical values.

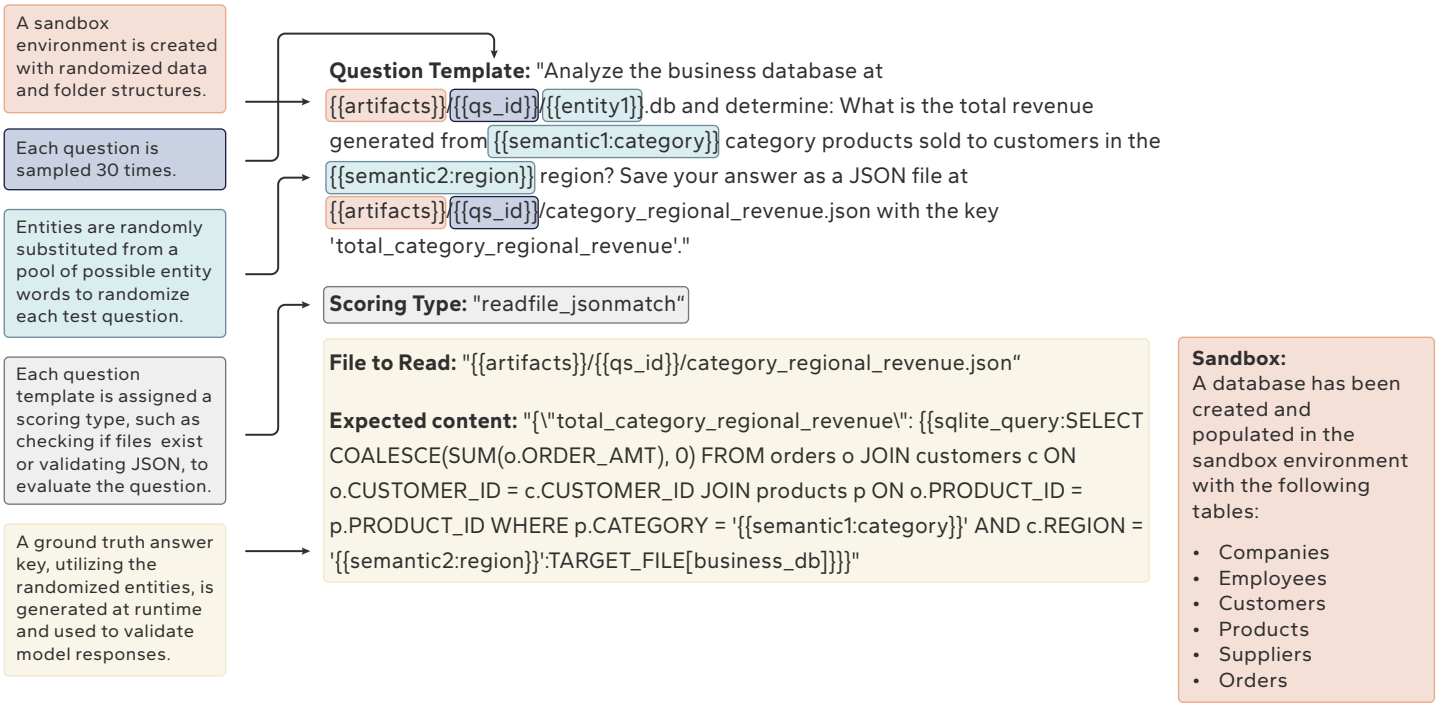


Figure 4: *Database Question Template Overview*

Example:

"Analyze the business database at test_artifacts/q503_s11/harbor.db and determine: What is the total revenue generated from technology category products sold to customers in the west region? Save your answer as a JSON file at test_artifacts/q503_s11/category_regional_revenue.json with the key 'total_category_regional_revenue'."

Correct Answer:

File:
test_artifacts/q503_s11/category_regional_revenue.json

Contents:
{"total_category_regional_revenue": 10000}

The correct answer correctly retrieved the answer to the query, formatted it as JSON, and created the output file in the correct directory.

Incorrect:

File:
test_artifacts/q503_s11/category_regional_revenue.json

Contents:
{"total_category_regional_revenue": 500}

Any answer that doesn't match the expected answer key is incorrect. In this example, the agent correctly created a JSON file with correct formatting, but incorrectly queried the database to calculate the value for 'total_category_regional_value'.

Figure 5: *Database Question Example*

Additional technical details and design principles of the KAMI benchmark are further outlined by Kamiwaza and can be found here. Beyond the initial findings, Signal65 and Kamiwaza plan for continued development of the KAMI benchmark, with future versions expanding to include more models and expanded testing abilities to further evaluate the agentic abilities and enterprise readiness of LLMs.

The KAMI benchmark introduces a unique new capability within Signal65's AI benchmarking and analysis portfolio. Designed to go beyond traditional test suites, KAMI provides a structured yet flexible framework for evaluating AI models, systems, and applications under realistic enterprise conditions. Signal65 will use KAMI as a foundation for ongoing model validation and end-to-end testing within its AI Lab. Through this approach, Signal65 will generate meaningful AI performance insights and advance industry understanding of how enterprise AI should be measured and compared.

Testing Overview

The first release of the KAMI Benchmark contains 19 distinct question templates, grouped into 7 specific categories. All questions were sampled 30 times for each run of the KAMI test suite to accommodate the variance of the randomized questions. In addition, for each model tested, the entire test suite was run multiple times and models were scored using their mean accuracy over all runs. An overview of the test questions can be seen in Figure 6 below.

Category	Performance
Basic Reasoning	Respond only with a specific word.
	Respond with multiple specified words in a specified order.
File System Operations	Create specific files in a specified directory.
	Create specific directory structures and include various files.
Text Search and Extraction	Find two specific lines from a file.
	Find several specific lines from an extended file.
	Retrieve two specific words from a text file.
	Retrieve several specific words from an extended text file.
CSV Processing	Create JSON summary of a CSV file.
	Analyze business data across multiple CSV files. Answer 6 specific questions.
	Analyze business data across multiple CSV Files. Single question.

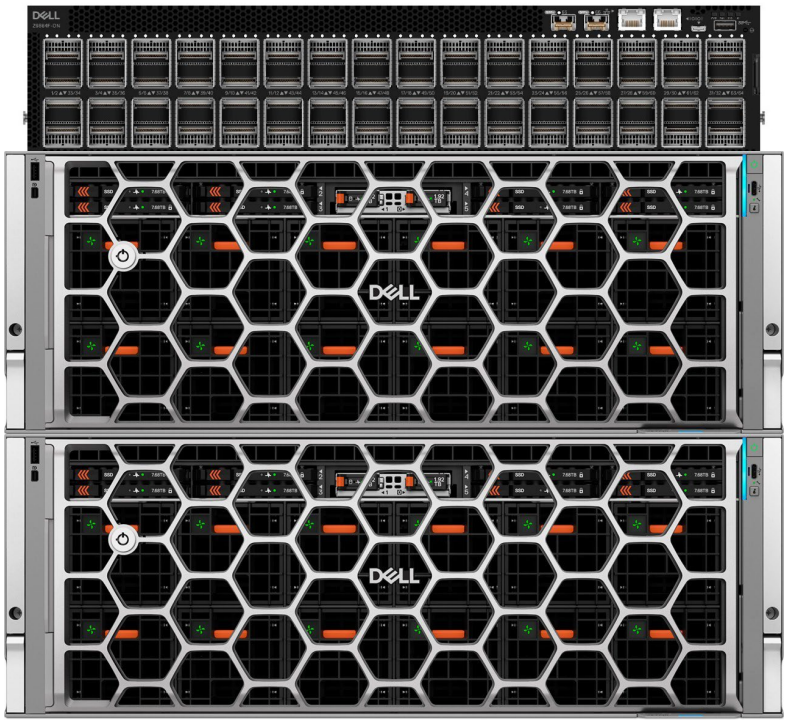
Category	Performance
Database Processing	Query business database to find number of orders over a specified value within a specified region.
	Analyze business database and create a comprehensive report. 6 specific questions.
	Analyze business database to find total revenue from a specified product in a specified region.
Database Processing (Guided)	Repeat simple database task with a hint given.
	Repeat complex database task with a hint given.
Response Format Instruction Following	Output answer to txt file.
	Output answer in JSON format.
	Output number only.

Figure 6: *KAMI Question Overview*

The initial KAMI benchmark was tested on 31 total models, resulting in over 170,000 total test conversations and over 5.5 billion tokens processed. Models were chosen to represent popular LLMs often considered in enterprise AI deployments. Models of various sizes and versions were additionally included to gain insight into their agentic capabilities. This iteration of the KAMI benchmark was primarily focused on open-source models, due to their ease of access and the wide range of models available. All open-source models were run on hardware in the Signal65 AI Lab, while proprietary models were run using their native API endpoints.

Example Testing Configuration

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 in a RoCEv2 fabric, these solutions achieve deterministic, low-jitter throughput across workloads. This balance of memory capacity, compute density, and network efficiency enables businesses to deploy Agentic AI systems for the most demanding reasoning workloads.



- Dell Z9864F-ON (Tomahawk 5)
- Enterprise SONiC 4.4.0
- 2 Dell PowerEdge XE7745:
- Dual AMD EPYC 9555
- 2.3 TB DDR5 Memory
- 8 NVIDIA RTX Pro 6000 Blackwell Server
- 8 Broadcom BCM57608 400GbE RoCEv2 Network Controller
- Ubuntu 24.04 LTS
- CUDA 13.0 / Driver 580.95.05
- vLLM v0.10.2

Figure 7: *Example Test Configuration*

Results

The full results of the KAMI v0.1 benchmark can be seen in Figure 8 below.

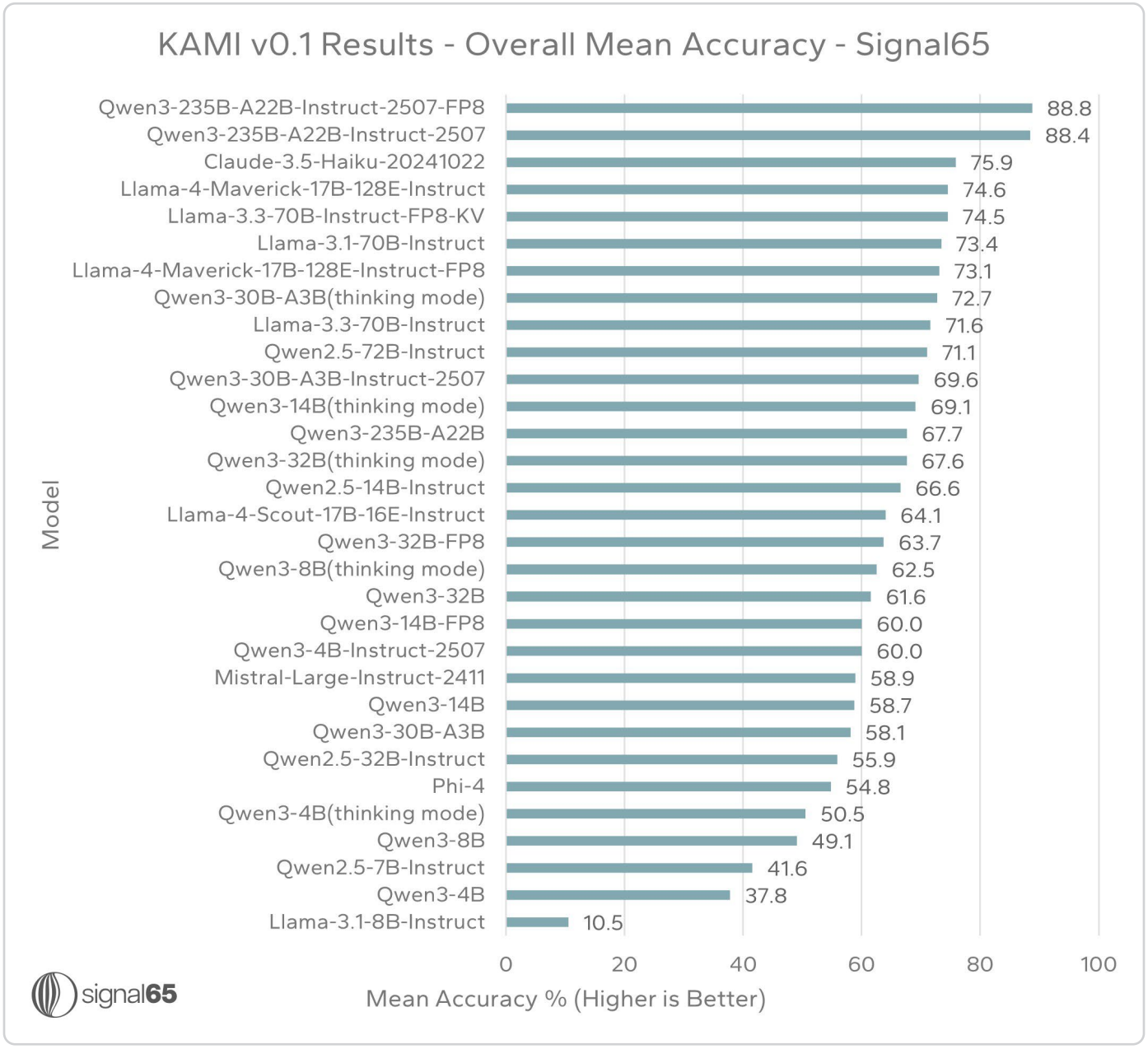


Figure 8: *KAMI v0.1 Results Overview*

Testing found Qwen3-235B-A22B-Instruct-2507-FP8 and Qwen3-235B-A22B-Instruct-2507 to achieve the highest overall scores with mean accuracies of 88.8% and 88.4% respectively. Notably, these two models were the only models to record an average accuracy over 80%.

The remainder of the top five performing models, Claude-3.5-Haiku-20241022, Llama-4-Maverick-17B-128E-Instruct, and Llama-3.3-70B-Instruct-FP8-KV, all obtained relatively competitive scores, ranging from 74.5% to 75.9% accuracy.

At the low end, Llama-3.1-8B scored the lowest overall accuracy at 10.5%. In addition, three other models also scored below 50% accuracy, including Qwen3-4B, Qwen2.5-7B-Instruct, and Qwen3-8B.

A greater understanding of each model's agentic performance can be gained from evaluating the results for each test category.

Model Size

A significant consideration for enterprises selecting LLMs is model size. In general, new model development has resulted in increasingly large models, with the general viewpoint being that larger models achieve better results. On the other hand, hardware limitations, and the associated cost considerations, may sway organizations to instead select smaller models. Additionally, ongoing model development and new architectures have led to the emergence of some small models that are thought to be highly competitive even with much larger LLMs.

By running the KAMI v0.1 benchmark across a wide range of LLMs, the results can be used to evaluate how models of various sizes achieve agentic tasks. To create such an evaluation, models have been grouped into four distinct groups:

- Small Models: < 10B Parameters
- Medium Models: 10B – 50B Parameters
- Large Models: 50B – 100B Parameters
- Very Large Models: > 100B Parameters

While there is still significant variation between models in each group, this rough grouping allows an overview of model performance based on size. It should be noted, that for this exercise, Mixture of Experts models were grouped by their total number of parameters rather than their active parameter counts. Additionally, Claude-3.5-Haiku-20241022 has been omitted from this analysis, as it does not have an officially documented parameter size.

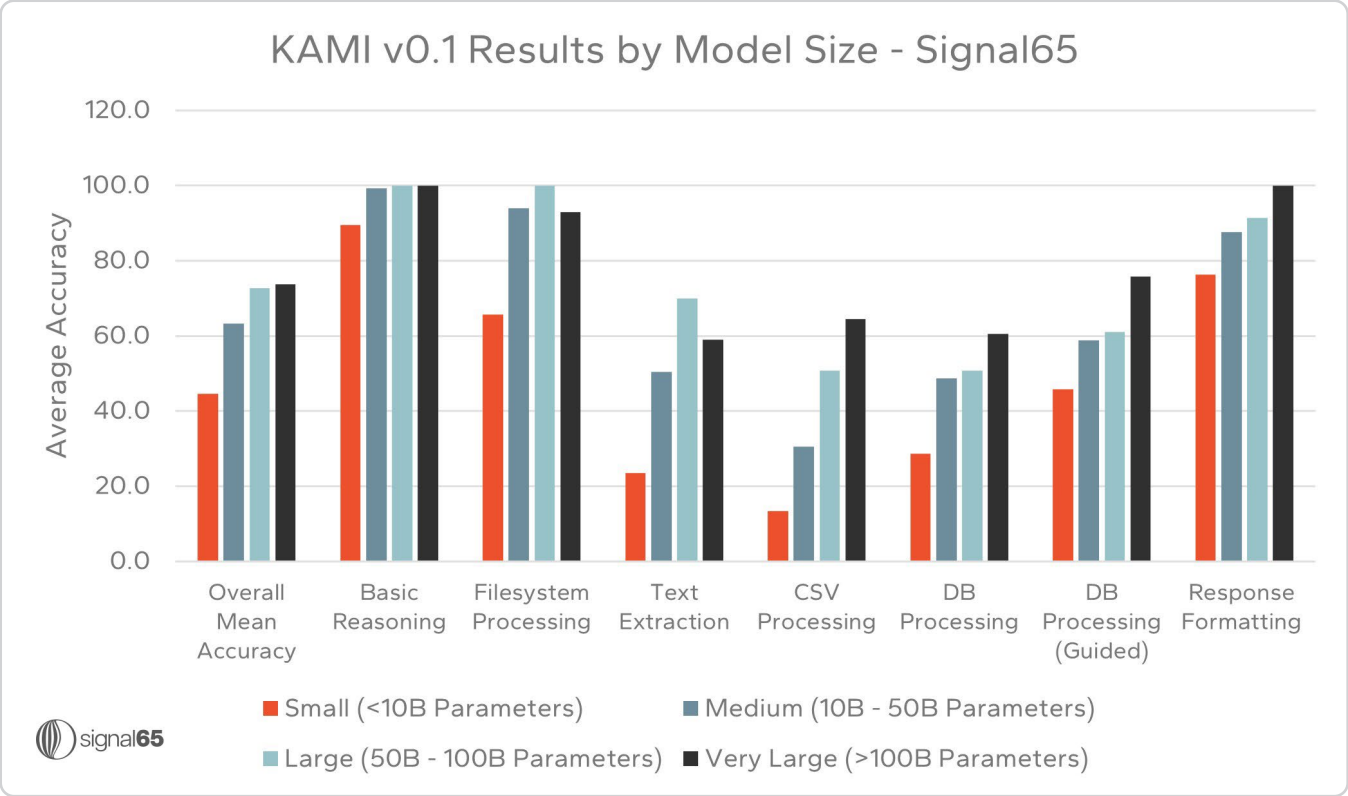


Figure 9: *KAMI v0.1 Results by Model Size*

The overall mean accuracy shows a clear trend that follows conventional thinking – on average, larger models perform better, with the large and very large model groupings far outperforming smaller models. The very large models (>100B parameters) show clear advantages across the CSV processing, database processing, and instruction following tasks, while the large (50B-100B parameters) group achieved the highest average scores for filesystem processing and text extraction. The medium models (10B-50B parameters), however, were found to remain competitive with the very large models for filesystem tasks and competitive with the large models for database processing and response formatting. The small models were found to be at a clear disadvantage across all test categories.

These results indicate that, in general, very small models may not be suitable choices for agentic AI applications. While they also indicate that the very large models will offer the greatest overall performance, the results for the medium and large models demonstrate that they may provide a reasonable choice for organizations seeking to balance hardware requirements, depending on the specific use case. Notably, the highest performing model in the medium category: Qwen3-30B-A3B (thinking mode), achieved an overall accuracy score of 72.7%, which was highly competitive with the top models in the large grouping, as well as some of the lower performing models in the very large grouping.

Quantization

Another aspect to evaluate within the KAMI v0.1 benchmark results is the impact of quantization on model performance. Quantization is typically utilized for efficiency; however, it comes with the risk of losing accuracy. Within the models selected for the KAMI v0.1 benchmark, several models were tested with both their full weights as well as a quantized fp8 version, enabling an evaluation of the quantization on agentic task performance.

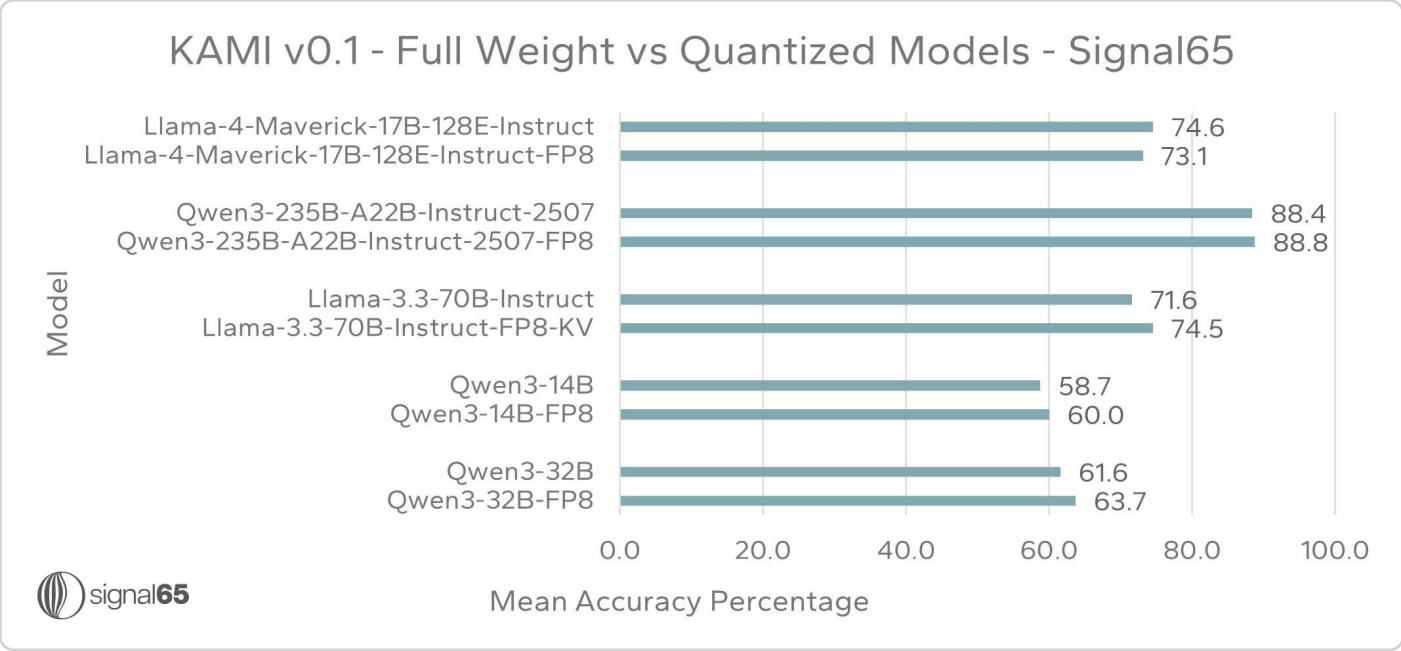


Figure 10: *KAMI v0.1 Full Weight vs Quantized Models*

On average, all of the thinking-enabled Qwen3 variations outperformed their non-thinking counterparts. The thinking variations achieved notable advantages across the Information Finding, CSV Processing, and non-hinted database tasks. Interestingly, the inclusion of hints enabled much more competitive performance by the non-thinking models.

In four of the five model pairs, the quantized version actually outperformed its full weight counterpart. In all cases, both models achieved very similar accuracies, and the differences may be a result of variation between test runs. While this testing does not provide enough to conclude that quantization actually improves model performance, it indicates that the loss of accuracy associated with fp8 quantization has minimal negative impact on completing agentic tasks.

Thinking vs Non-Thinking Models

In addition to including both quantized and non-quantized models for comparison, the KAMI benchmark included several Qwen3 models that were run in both thinking and non-thinking modes, which enables a comparison of the impact of thinking on otherwise identical models.

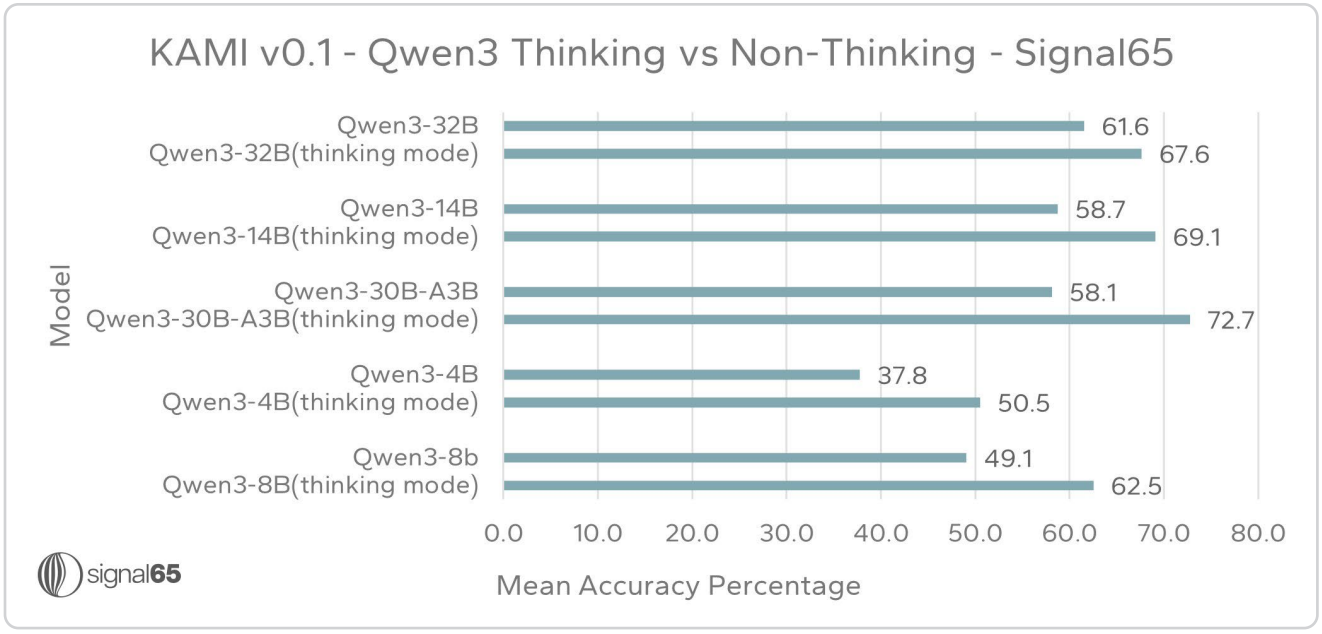


Figure 11: *Thinking vs Non-thinking Models Overview*

Conclusion

Signal65 analysis of the initial KAMI benchmark results serve as tools to evaluate how LLMs perform in realistic agentic scenarios. KAMI differentiates itself from other popular LLM benchmarking approaches by utilizing dynamic question generation and a focus on real-world agentic tasks. For enterprise organizations building agentic AI applications, this approach offers valuable insight into model performance and can assist organizations in selecting the right models to build successful AI agents.

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**

Understanding the agentic capability of models is important for enterprise organizations when considering multi-agent applications, where different models may be required to maximize the potential of each agent. Selecting accurate models for each agent role becomes additionally important as a single inaccurate agent can impact the entire application, complicating the model selection processes. Using tools like the KAMI benchmark can provide context and data to help understand the strengths and weaknesses of various LLMs in real agentic scenarios. Additional details on the KAMI benchmark design and initial runs may be found on the Signal65 site.

For business leaders navigating emerging AI opportunities, investing in infrastructure that scales with evolving needs is crucial for driving efficiency, innovation, and competitive edge. The Dell PowerEdge XE7745 offers a robust foundation for deploying and expanding Agentic AI, ensuring data sovereignty and best in class open solutions. Signal65 testing underscores the important role of purpose-built AI hardware in optimizing performance.

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 Thor2 NICs, and the sophistication of evolving models and agentic designs, will be best positioned to harness the full potential of AI-driven knowledge systems, driving innovation and maintaining competitive advantage in an increasingly AI-powered business landscape.

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.

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