# Evaluating Agentic AI

## Dell AI Solutions with Open Models

As AI shifts from experimental concept to mission-critical capability, Agentic AI will define new enterprise AI applications. The Kamiwaza Agentic Merit Index (KAMI) dynamic benchmark evaluates practical agentic task performance across LLMs.

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 builds uniquely randomized test environments for every run to produce contamination-free execution-focused evaluation.

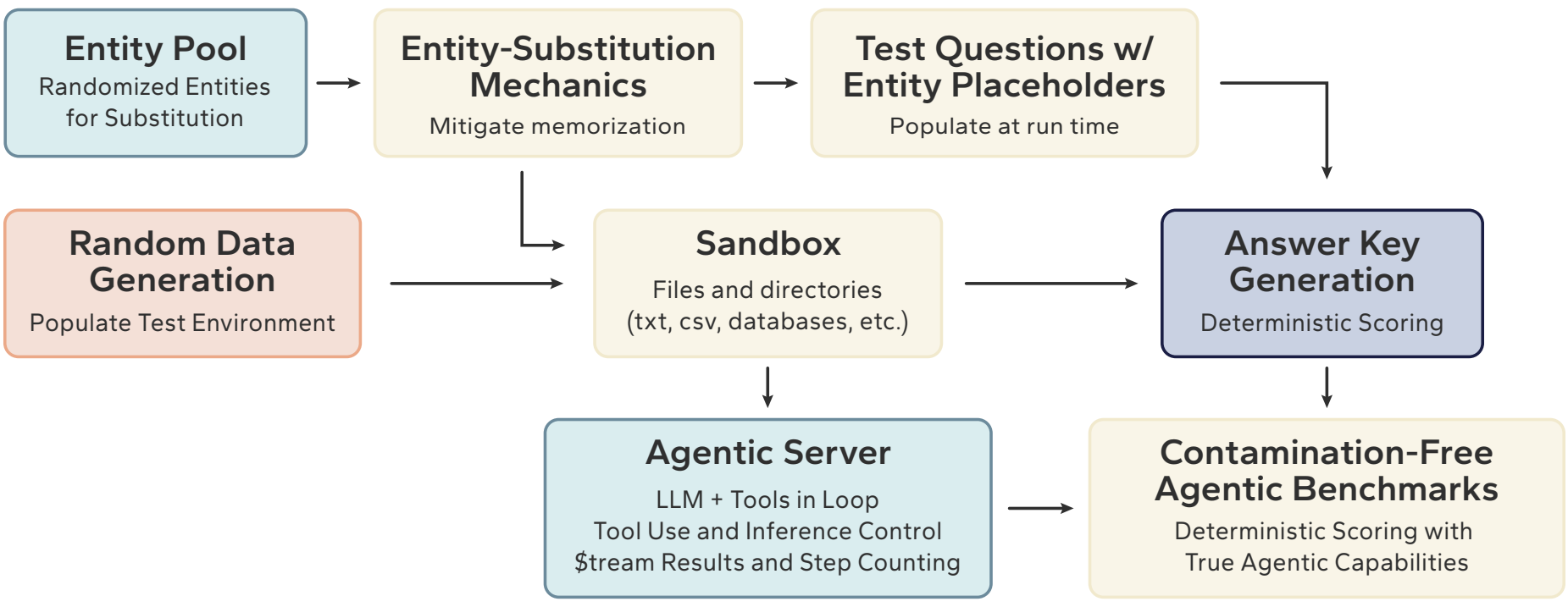
### Elements

- Sandboxed file and database environments
- Dynamic substitution for unique instructions
- Runtime-generated ground-truth answer keys
- Tool-calling loops through an agentic server

### 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

### PICARD Framework




### 1 Model Size Strongly Correlates with Agentic Accuracy

- **Very large (>100B)** models performed best
- **Large (50–100B)** models excel at text extraction
- **Medium (10–50B)** models surprisingly competitive
- **Small (<10B)** models performed poorly

### 2 FP8 Quantization had Minimal or No Negative Impact

### 3 Thinking Models Outperformed Non-Thinking Counterparts

| 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 % |

***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 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.

For more, see the full report on the Signal65 website.



In partnership with: KAMIWAZA, DELL Technologies