signal65

September 2026
Kamiwaza · Dell Technologies · Broadcom

Trusted Answers at Enterprise Scale

Retrieval reliability for enterprise AI, measured on a Dell and Broadcom stack

Enterprise AI's most common job is answering questions from the organization's own documents: contracts, policies, operational reports, and records. The business case rests on one question — **can the answers be trusted?**

THE RETRIEVAL TRUST GAP

Retrieval quality holds up at moderate document scale and thins out sharply as corpora grow. Deployment configuration moves the number as much as model choice does.

27 of 91

BROAD CHOICE AT 32K

models exceeded 95% accuracy at moderate workloads — a wide field of viable candidates.

3 of 54

SCARCITY AT 200K

models stayed above 95% accuracy at larger workloads — the field collapses at production scale.

up to **64%**

CONFIGURATION AS A LEVER

relative accuracy gain from enabling reasoning — one model moved from 58.8% to 96.6%.

HOW RIKER WORKS

Traditional benchmarks start from existing documents, then annotate answers by hand — slow, hard to scale, and vulnerable to training-data contamination. RIKER inverts the process: it generates documents from predefined ground-truth answers, so grading stays deterministic and every run is fresh.

Ground Truth Database → Defines entities and relationships

Coherent Simulated Universe → Consistent relationships across documents

Template Document Generation → Synthetic corpus with deterministic answer keys

Scalable, Reliable Benchmark → Contamination-resistant, no human annotation

REGENERABLE CORPORA — fresh instances generated at will, preventing memorization and contamination

Govern Demand
4 TASKS
Direct, indirect, conditional, and complex extraction within one source document.

Multi-Document Aggregation
6 TASKS
Counting, summation, comparison, enumeration, multi-hop, and temporal reasoning across sources.

Hallucination Probing
2 TASKS
Nonexistent entities and absent fields — does the model say "unknown" when it should?

THE FIELD NARROWS AS DOCUMENTS GROW

Testing covered 91 models at 32K context, 72 at 128K, and 54 at 200K, using corpora of commercial leases, HR records, and facility field reports. Advertised context windows did not predict which models held their accuracy.

CONTEXT	TESTED	ABOVE 90% ACCURACY	ABOVE 95%
32K	91	45	27
128K	72	24	10
200K	54	15	3

At 200K context, only **three of 54 models** sustained the 95% standard. A shortlist built on 32K results alone will not survive contact with production-scale corpora.

OVERALL MEAN ACCURACY — 200K CONTEXT

1	Qwen3.5-397B-A17B (Thinking)	95.63%
2	Qwen3.5-397B-A17B-KV-FP8 (Thinking)	95.54%
3	GPT-5.4 (Medium Reasoning)	95.25%
4	Qwen3.5-122B-A10B (Thinking)	94.53%
5	Kimi-K2.5 (Thinking)	94.39%
6	Qwen3.5-122B-A10B-KV-FP8 (Thinking)	93.94%
7	Qwen3.5-27B (Thinking)	93.44%
8	Qwen3.5-397B-A17B	93.34%

WHERE ACCURACY BREAKS DOWN

Averaged across all evaluated models, relative to 32K performance. Cross-document work degrades fastest; knowing when an answer is absent proves the most durable capability.

−25.6%
Multi-document aggregation accuracy at 200K — the clearest production bottleneck

2×
Faster degradation for aggregation than for single-document extraction

−9.8%
Hallucination-probe accuracy at 200K — the most stable of the three categories

+64.3%
Largest relative gain from enabling reasoning at 32K context

THE TESTED OPEN-MODEL STACK

Open models ran in the Signal65 AI Lab on an on-premises stack, keeping sensitive documents under the organization's control. Proprietary models were run through their native API endpoints.

DELL POWERSWITCH Z9864F-ON + 2× DELL POWEREDGE XE7745

Dell PowerSwitch Z9864F-ON — The Fabric

Built on Broadcom Tomahawk 5 silicon and running Enterprise SONiC 4.4.0, the switch interconnects the cluster over 400GbE RoCEv2. Long-context inference raises memory demand for the KV cache, and distributed serving adds inter-node communication — the fabric is what keeps that traffic predictable.

Dell PowerEdge XE7745 — The Compute

A high-memory, multi-GPU platform for large-context inference and high-volume retrieval. Testing also used PowerEdge XE9680 with NVIDIA H200 and AMD MI300X. Each XE7745 node in the tested cluster:

- Dual AMD EPYC 9555 · 2.3 TB DDR5 memory
- 8× NVIDIA RTX Pro 6000 Blackwell Server
- 8× Broadcom BCM57608 400GbE RoCEv2 network controllers
- Ubuntu 24.04 LTS · CUDA 13.0 / Driver 580.95.05 · vLLM v0.10.2

FROM RISK TO RESOLUTION

The findings map directly onto the exposures an enterprise retrieval program places on the budget owner's desk.

Business Risk	What the Testing Shows
Procurement Risk	Advertised context windows are capacity limits, not evidence of retrieval quality. Specify accuracy at corpus scale in evaluation criteria.
Pilot-to-Production Risk	Single-document pilots overstate readiness; cross-document aggregation fell 25.6% on average at 200K context.
Agentic Program Risk	Per-step errors compound: five retrieval-dependent steps at 90% reliability complete cleanly only about 59% of the time.
Compliance and Audit Risk	Recognizing absent information is a separate capability from retrieval, and should be measured on its own.

Validate Retrieval Before Delegating Autonomy

Agents act on what they retrieve, so a retrieval error becomes an action error. Evaluate models, configurations, and infrastructure together against the workloads the organization actually intends to run — not against specifications alone.

In partnership with: KAMIWAZA | DELL Technologies | BROADCOM