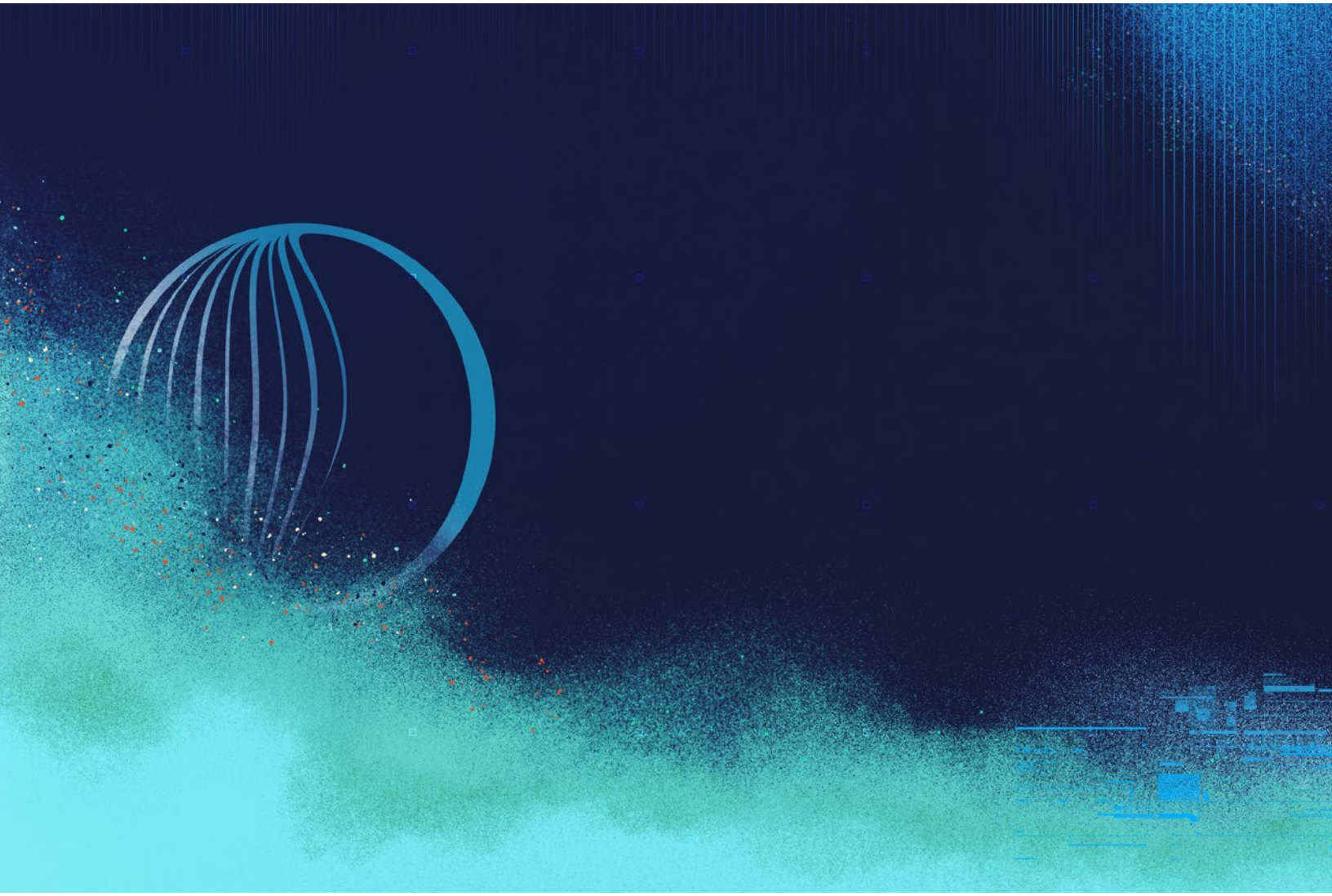
signal65



EXECUTIVE SUMMARY

Trusted Answers at Enterprise Scale

Dell AI Solutions with Open Models

AUTHORS

Mitch Lewis
Performance Analyst | Signal65

Brian Martin
AI Data Center Performance | Signal65

SEPTEMBER 2026

IN PARTNERSHIP WITH

KAMIWAZA

DELL Technologies

BROADCOM

Overview

Enterprise AI has moved from pilots to production, where its most common job is answering questions from an organization's documents: contracts, policies, operational reports, and records. Business results including faster decisions, lower operating costs, and institutional knowledge on demand rest on a deceptively simple question: **can the answers be trusted?**

A system that retrieves the wrong clause, mis-totals figures across reports, or fabricates answers creates operational, financial, and compliance exposure. The stakes rise with agentic AI: agents retrieve at nearly every step, per-step error rates compound across a workflow, and a fabricated answer becomes a fabricated action.

Signal65 and Kamiwaza built a benchmark specifically to answer that question with evidence, evaluating 91 models on the retrieval work enterprises actually run. Testing across corpora from 32K to 200K tokens ran in the Signal65 AI Lab on Dell PowerEdge compute with Broadcom-based 400GbE networking.

Key Highlights

Broad Choice at Moderate Scale

27 of 91 models exceeded 95% accuracy at moderate (32K) workloads

Scarcity at Production Scale

3 of 54 models remained above 95% accuracy at larger (200K) workloads

Configuration as a Cost Lever

Reasoning improved retrieval accuracy by **up to 64%**

For a buyer, the decision reduces to two variables: **how much accuracy a model holds at the organization's actual document scale, and how much it loses as that scale grows.** Advertised context windows measure neither.

Two Questions Behind the Deployment Decision

Scale. Single-document lookup and cross-document aggregation are different workloads with different economics. At 32K context a broad field performs well, but at 200K, aggregation accuracy fell by an average of 25.6%, more than twice the rate of single-document tasks. Because most pilots are built on simple lookups, they can materially overstate production readiness, and the shortfall surfaces only after the budget is committed.

Trust. Whether a model recognizes an answer is absent is partly independent of whether it can find one. The strongest models reliably acknowledged missing information and stayed comparatively stable as context grew, which is what keeps a retrieval error from becoming a confident fabrication. Configuration is a lever on both: enabling reasoning moved one model from 58.8% to 96.6% accuracy at 32K, a 64.3% relative gain, and belongs in the evaluation alongside latency, throughput, and cost.

Benchmarked at Production Scale

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. At 32K, roughly the scale of a contract review or policy lookup, the top 30 models all exceeded 90% accuracy, 27 reached at least 95%, and five surpassed 98%. Qwen3.5-397B-A17B (Thinking) led at 99.1%.

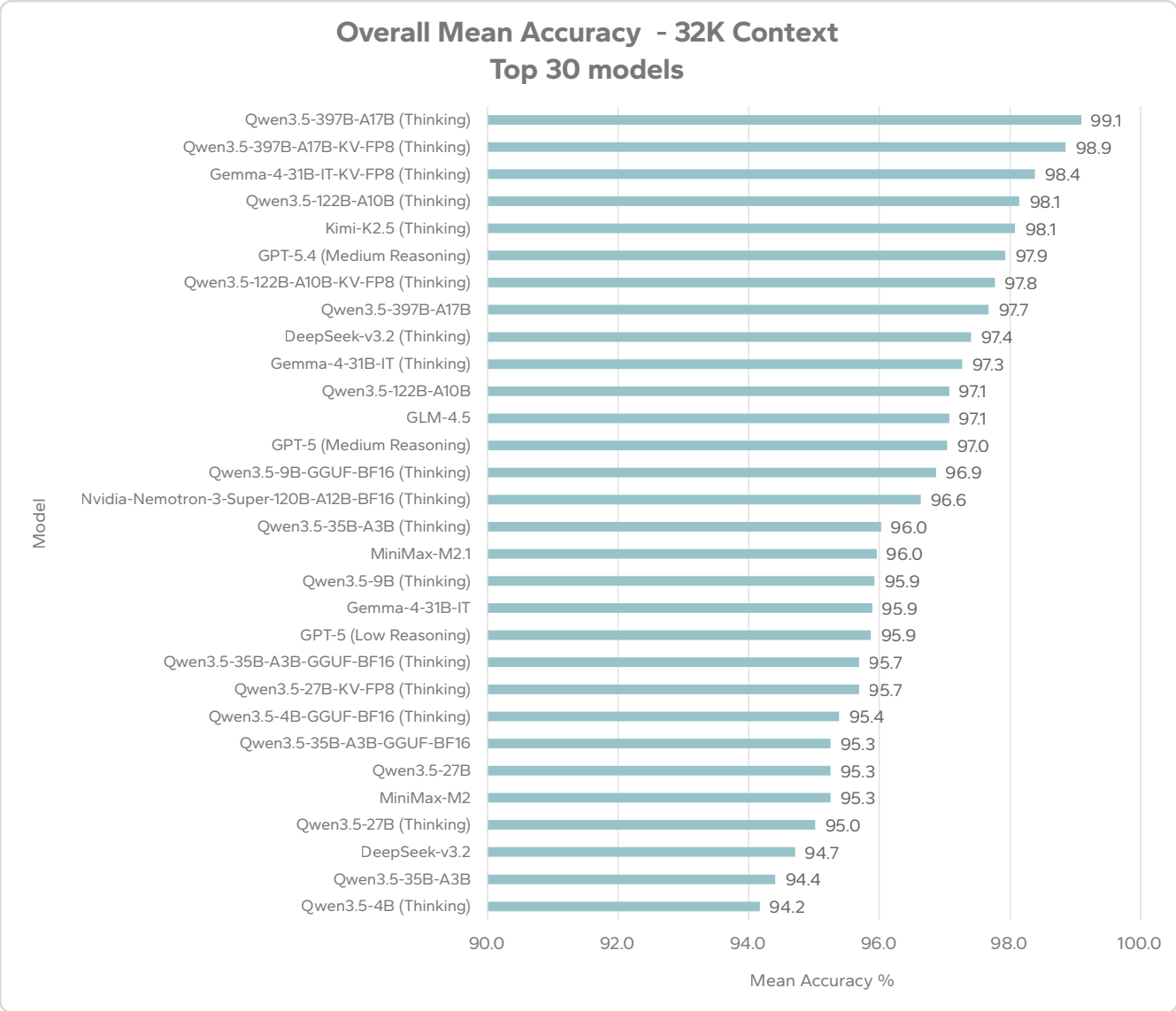


Figure 1: *Overall Mean Accuracy – 32K Context Size*

That breadth narrows sharply as documents grow: 24 of 72 models exceeded 90% at 128K, and only 15 of 54 at 200K, where just three cleared 95%. A shortlist built on 32K results alone will not survive contact with production-scale corpora.

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.

The Bottom Line

Trusted retrieval is becoming a prerequisite for enterprise AI, not a differentiating feature. A small set of models sustained strong retrieval and abstention at every workload size tested, and the gap between those leaders and the broader field makes model and configuration selection a decision with measurable operational consequences. Evaluate models, configurations, and infrastructure together against the workloads the organization actually intends to run, not against specifications alone.

Excerpted from Trusted Answers at Enterprise Scale, Signal65, August 2026. Complete per-model results are available from Signal65.

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 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

ACKNOWLEDGEMENTS

The authors would like to thank Kamiwaza for the RIKER benchmark and Dell Technologies for providing access to the PowerEdge XE9680 and XE7745 hardware platforms 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