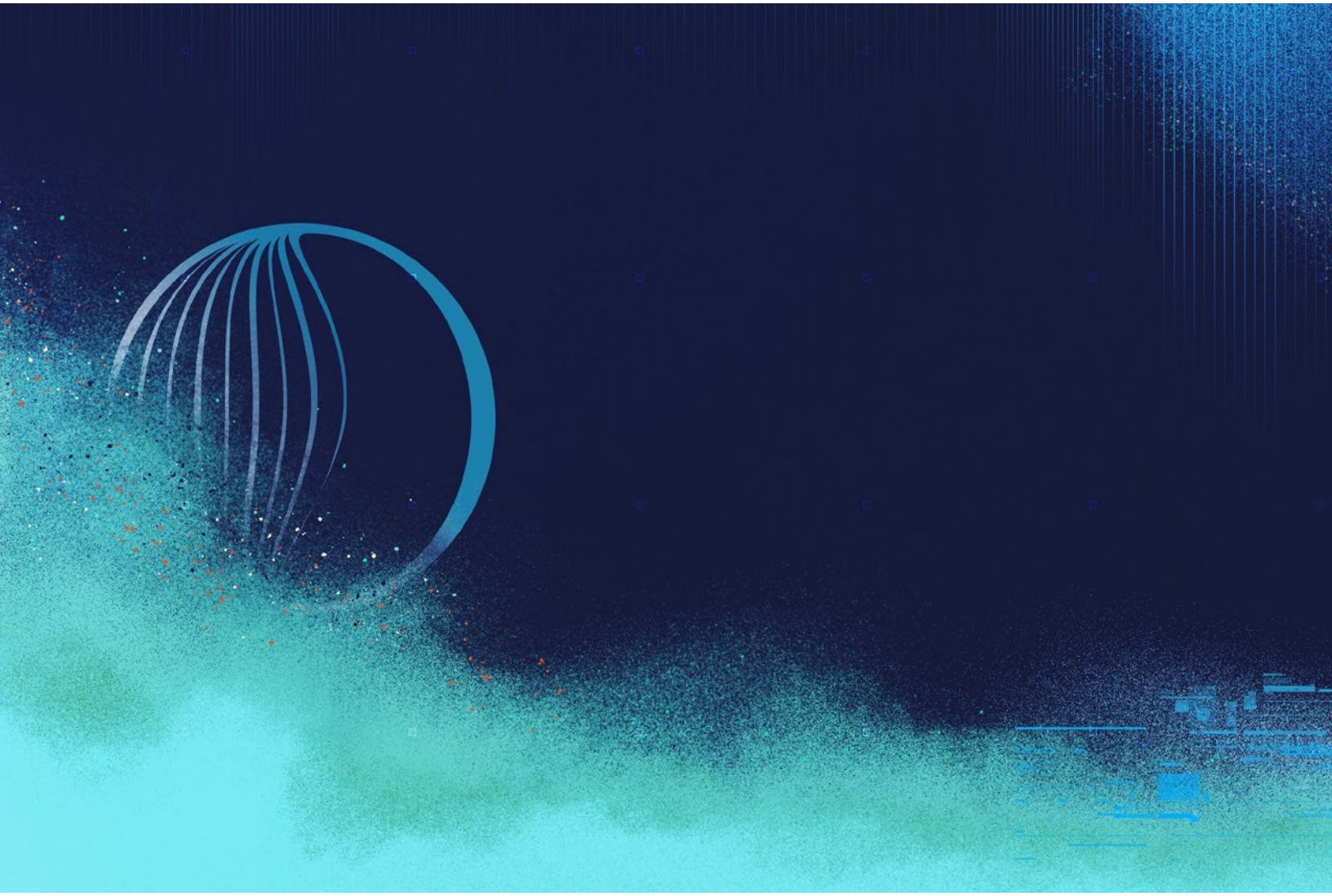
signal65



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

Executive Summary

Enterprise AI has moved from pilots to production workflows, where its most common job is answering questions from the organization's own documents: contracts, policies, operational reports, and records. The business value of these systems, from faster decisions and lower operating costs to institutional knowledge available on demand, depends on a deceptively simple question: can the answers be trusted? An AI system that retrieves the wrong clause from a contract, mis-totals figures across reports, or fabricates an answer when none exists creates operational, financial, and compliance risk, and undermines the confidence needed for adoption to spread.

Key Highlights

27 of 91 models achieved at least 95% accuracy with 32K-token document inputs.

3 of 54 models exceeded 95% accuracy with 200K-token document inputs.

At 32K, reasoning raised one model's accuracy **from 58.8% to 96.6%.**

The stakes are rising as enterprises move toward agentic AI, systems that do not simply answer questions but carry out multi-step business processes. Agents retrieve information at nearly every step of their work: checking a policy, querying records, reconciling figures before acting. Retrieval reliability is therefore not one AI capability among many; it is the foundation on which agentic initiatives stand or fail, because small per-step error rates compound across a workflow, and a fabricated answer becomes a fabricated action.

Signal65 and Kamiwaza developed RIKER (Retrieval Intelligence and Knowledge Extraction Rating) to compare how accurately models find facts, combine information across documents, and recognize missing information within supplied document collections. This report helps enterprise teams shortlist models and configurations for further testing. It evaluates model behavior within provided context windows, not complete retrieval-augmented generation (RAG) pipelines or autonomous agent workflows. Testing used synthetic commercial leases, HR records, and facility field reports, with document inputs up to 200K tokens—roughly 150,000 words.

Open models make their model weights available for deployment under their respective licenses, allowing organizations to run them on infrastructure they control. Signal65 conducted the open-model evaluation in its AI Lab using RIKER, developed with Kamiwaza. Dell Technologies provided access to PowerEdge hardware and technical expertise; Broadcom contributed network optimization collaboration and the Ethernet technology used in the tested stack. Proprietary models were evaluated through their native APIs.

What the testing found, and why it matters to the business:

- **High accuracy over large document inputs is achievable, but uncommon:** Only three of 54 models exceeded 95% overall accuracy at 200K tokens. These controlled results support model selection; they do not establish the accuracy of a complete enterprise retrieval system.
- **Advertised context windows are not a procurement metric:** Models with similar advertised limits diverged sharply as workloads grew. Evaluate retrieval stability at the document scale the organization actually operates, not from token limits alone.

- **Cross-document work is the key bottleneck:** At 200K context, multi-document aggregation accuracy declined by an average of 25.6% relative to 32K performance, more than twice the rate of single-document tasks. Pilots built on simple lookups can overstate production readiness.
- **Retrieval reliability determines agentic readiness:** Because agents retrieve information throughout multi-step workflows, per-step accuracy compounds. Assuming independent failures and five retrieval-dependent steps, 90% reliability at each step yields about a 59% probability of completing the workflow without a retrieval error. Validate retrieval stability before delegating consequential actions.
- **Deployment configuration is a business lever, not an IT detail:** At 32K context, enabling reasoning ("thinking") improved one model from 58.8% to 96.6% accuracy, a gain of 37.8 percentage points, or 64.3% relative. Test reasoning modes independently; the effect varied by model and context size.
- **Model selection can materially reduce hallucination risk:** The best models reliably acknowledged when information was absent, and that behavior remained comparatively stable as context grew. "Can the model find the answer?" and "does it know when the answer is absent?" should be evaluated separately.
- **Evaluate model, configuration, and infrastructure together:** In the tested environment, Dell PowerEdge compute and a Broadcom-based 400GbE RoCEv2 fabric formed the on-premises stack used for open-model evaluation. The benchmark demonstrates that the workloads ran on this configuration; it does not isolate the performance contribution of individual infrastructure components.

Knowledge Retrieval for Enterprise AI

Enterprise AI systems increasingly rely on retrieval-based workflows to access organizational knowledge distributed across documents, databases, and internal systems. These systems provide information to LLMs in a variety of ways, ranging from documents supplied directly in a chat session to retrieval-augmented generation (RAG) pipelines, knowledge graphs, and agentic retrieval systems. These workflows also introduce challenges involving retrieval accuracy, cross-document reasoning, and hallucination.

Despite the growing importance of enterprise knowledge retrieval, assessing model performance for these tasks has proven difficult. Many existing benchmarks fail to accurately represent real-world enterprise retrieval tasks, instead focusing primarily on surface-level extraction or pattern matching. Benchmarks that rely on static datasets become vulnerable to contamination through model training. Other approaches depend on LLM-as-a-judge evaluation methodologies, introducing additional uncertainty and bias into scoring.

These limitations create a gap in understanding model performance for one of the most common enterprise AI workloads: retrieval and reasoning over organizational knowledge. Previously, Signal65 and Kamiwaza collaborated to develop the KAMI benchmark, which evaluates agentic AI capability using a dynamically generated benchmark framework. Building on this approach, Signal65 and Kamiwaza developed RIKER (Retrieval Intelligence and Knowledge Extraction Rating), a benchmark designed to evaluate enterprise knowledge retrieval and understanding. Together, they provide complementary evaluation frameworks for measuring AI performance across enterprise retrieval and agentic workloads.

Introducing RIKER

RIKER (Retrieval Intelligence and Knowledge Extraction Rating) is a benchmark designed to evaluate enterprise-oriented retrieval and contextual understanding tasks. Unlike traditional knowledge benchmarks that primarily measure memorization or surface-level extraction, it evaluates a model's ability to retrieve, aggregate, and reason over information distributed across realistic document corpora (multiple collections of related documents).

The benchmark differentiates itself from many existing approaches by combining dynamically generated test environments with deterministic grading. This approach reduces susceptibility to benchmark memorization while avoiding the uncertainty and bias introduced by LLM-as-a-judge evaluation methodologies. To achieve this, it utilizes an inverted generation approach.

Traditional benchmark generation typically begins with pre-generated documents from which answers are later extracted and manually annotated. While effective for small-scale benchmarks, this approach often produces static datasets that are difficult to scale and increasingly vulnerable to contamination through model training. Alternative approaches that rely on LLM judges introduce additional variability and scoring uncertainty.

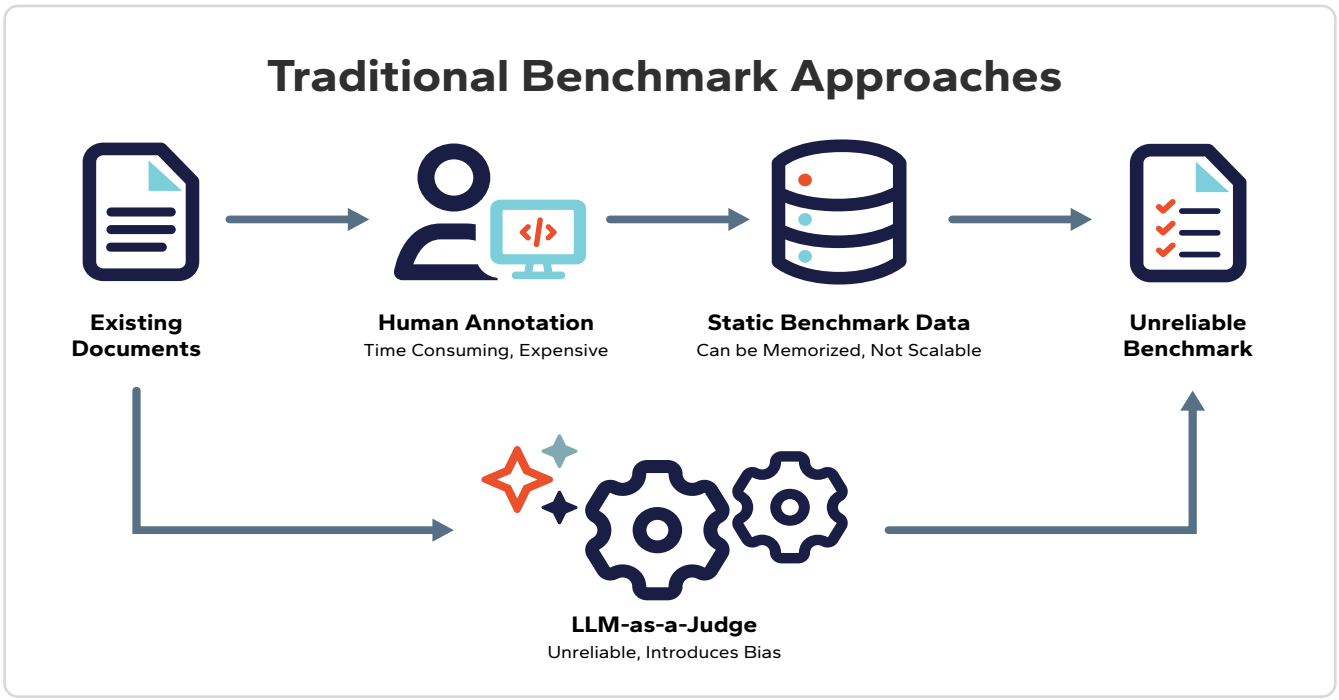


Figure 1: *Traditional Benchmark Approaches*

The benchmark instead reverses the dataset generation process by generating documents from predefined ground-truth answers. A structured database of entities and answers is used to populate randomized document templates, enabling dynamically generated document corpora while preserving deterministic evaluation. This approach allows it to generate unique test instances that are resistant to memorization while maintaining scalable, repeatable scoring.

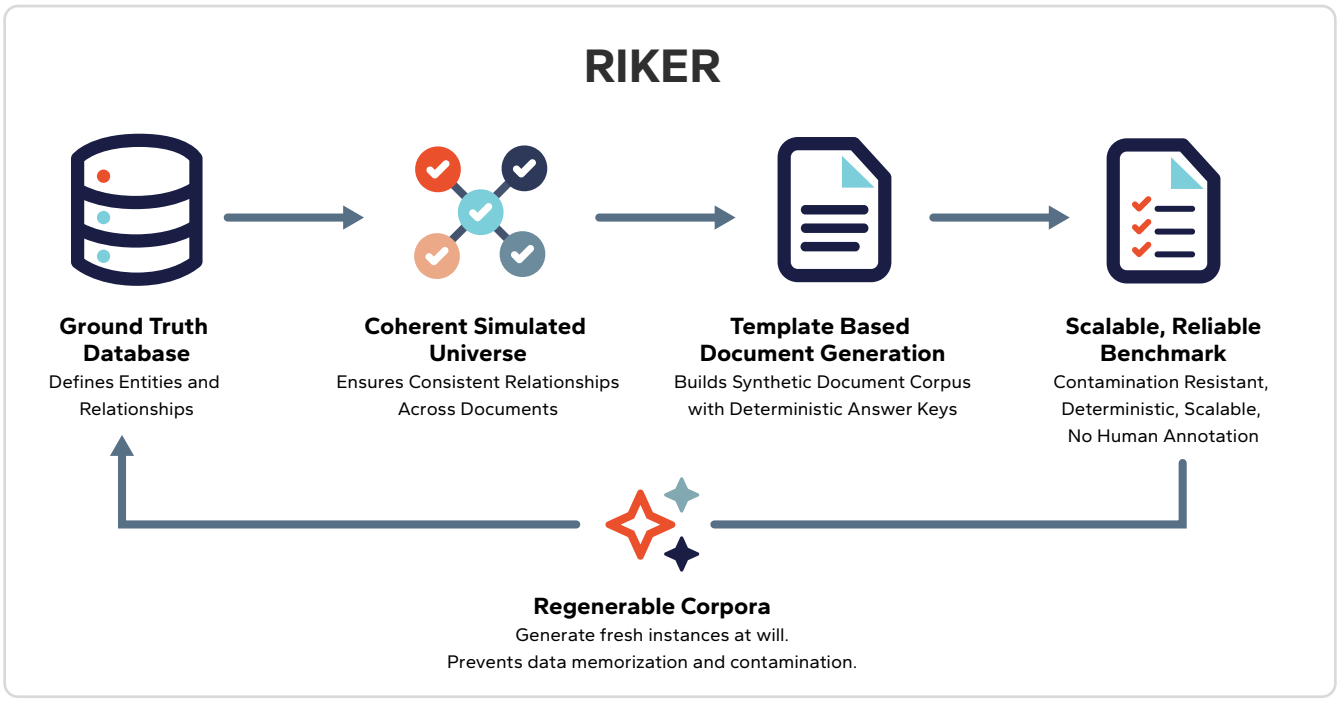


Figure 2: *RIKER Overview*

A key feature is the "Coherent Simulated Universe" methodology. Traditional synthetic datasets often generate documents independently, resulting in unrealistic inconsistencies across related records. For example, an HR manager referenced in one document may be associated with a different department in another because entities were populated independently during generation. The benchmark instead maintains consistent relationships between entities across documents, enabling coherent multi-document retrieval and aggregation tasks that better approximate enterprise knowledge environments.

The current release includes retrieval tasks spanning multiple enterprise-focused domains, including commercial leases, facility field reports, and HR records. Additional technical details about the benchmark and the coherent simulated universe are available from Kamiwaza.

Test Methodology

The RIKER benchmark suite currently includes 12 evaluation prompts spanning three retrieval-oriented task categories. Tasks range from direct extraction within a single document to multi-document aggregation and temporal reasoning. It also includes hallucination probes that evaluate whether models generate information absent from the provided corpus.

The current benchmark corpus includes documents spanning three enterprise-focused domains: commercial leases, facility field reports, and HR records.

Section	Q#	Test	Description	Example
Single Document Retrieval Tasks				
	1	Direct Extraction	Surface-level facts stated explicitly	"What is the monthly rent?"
	2	Indirect Extraction	Facts requiring minimal inference	"What is the lease duration?" (when start/end dates are given)
	3	Conditional Extraction	Facts from optional document sections	"What is the pet deposit?" (may be N/A)
	4	Complex Extraction	Facts requiring multiple conditions or cross-referencing within a document	"Who is the agent for Lessor X's lease with Lessee Y starting on a specified date?"
Multi-Document Aggregation Tasks				
	5	Counting	Totaling occurrences across multiple documents	"How many leases does Lessor X have?"
	6	Summation/ Averaging	Calculating summation or averages across multiple documents	"What is the total monthly rent across all leases?"
	7	Comparison	Comparing distinct values across multiple documents	"Which lessor has more leases, X or Y?"
	8	Enumeration	Find all instances related to a specific entity	"List all lessees for Lessor X."
	9	Multi-hop	Extract information across multiple examples	"What is Lessor X's most recent lease end date?"
	10	Temporal	Retrieve information across a specified period of time	"How many leases were active in Q3 2024?"
Hallucination Probing Tasks				
	11	Non-existent Entities	Questions about entities that do not appear anywhere in the corpus	"What is the monthly rent for Lessor X's Lease with Lessee Y starting on a specified date?" (These Lessors do not exist)
	12	Absent Information	Questions about optional fields that are absent from specific documents	"What is the early termination fee for Lessor X's lease?" (Early Termination fee is not specified)

Figure 3: *RIKER Benchmark Tasks*

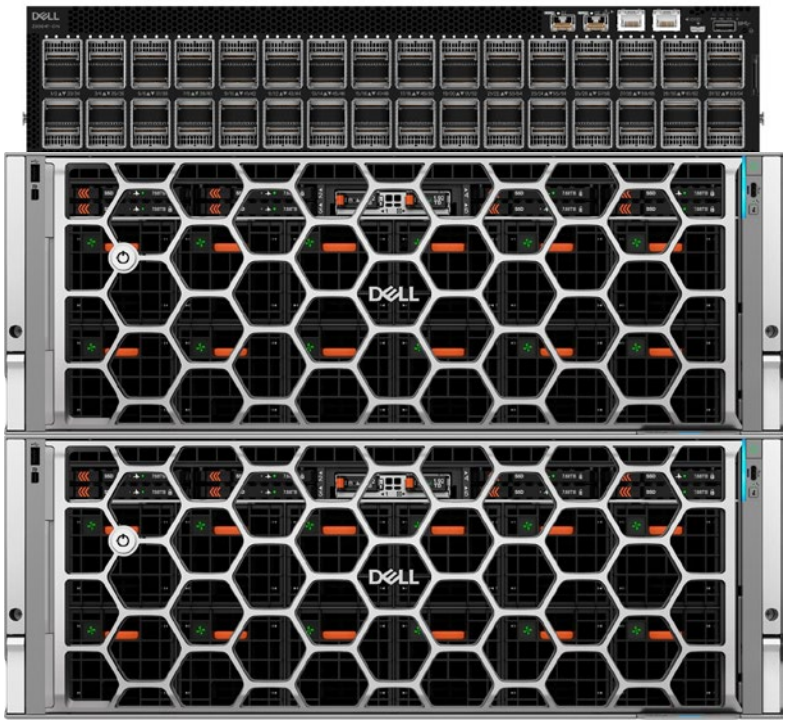
To reduce variance, each benchmark configuration was executed multiple times and averaged across runs. A context window is the model's capacity for processing tokens, the units into which text is divided. Here, context size refers to the amount of document text supplied for testing: 32K, 128K, or 200K tokens. In total, testing included 91 models at 32K context, 72 models at 128K context, and 54 models at 200K context.

Models were selected to represent LLMs commonly considered for enterprise deployment, with an emphasis on open models. Open models were run in the Signal65 AI Lab on the Dell PowerEdge and Broadcom-based Ethernet stack described below; proprietary models were run through their native API endpoints.

Open-Model Test Infrastructure

The open-model tests used Dell PowerEdge XE9680 with NVIDIA H200 and AMD MI300x GPU and Dell PowerEdge XE7745 with NVIDIA RTX Pro 6000 GPU. Broadcom BCM57608-based 400GbE adapters connected the nodes through a Dell Z9864F-ON switch using RoCEv2.

Together, Dell compute infrastructure and Broadcom-based Ethernet fabric form the tested on-premises infrastructure stack. Long-context inference increases memory demand for the key-value (KV) cache, and distributed serving adds inter-node communication. The BCM57608 adapters provide the scale-out connectivity layer, supporting 400GbE and RoCEv2 between inference nodes. This benchmark validated the combined environment but did not compare alternate fabrics or quantify the network's independent effect on accuracy, latency, or throughput.



- 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 4: *Open-Model Test Infrastructure*

Findings

Overall

Across all evaluated models, retrieval accuracy generally declined as context length increased. More notably, model performance divergence expanded substantially at larger context sizes, suggesting that advertised context window size alone is not a reliable indicator of effective enterprise retrieval capability.

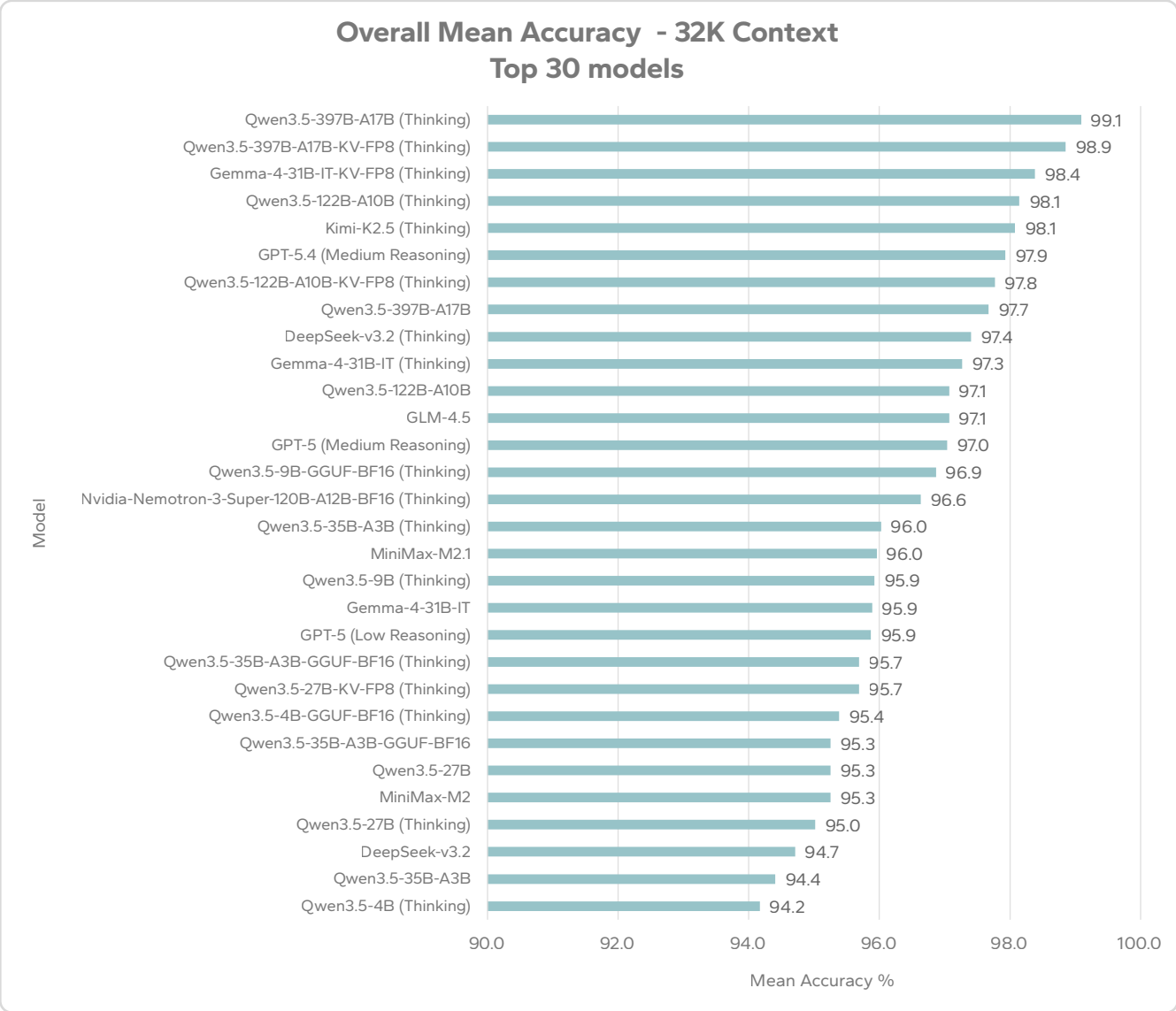


Figure 5: *Overall Mean Accuracy – 32K Context Size*

At 32K context, 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%, followed by its FP8-quantized variant at 98.9%. Overall, 45 of 91 models achieved at least 90%, indicating that many models can handle moderate-scale retrieval workloads.

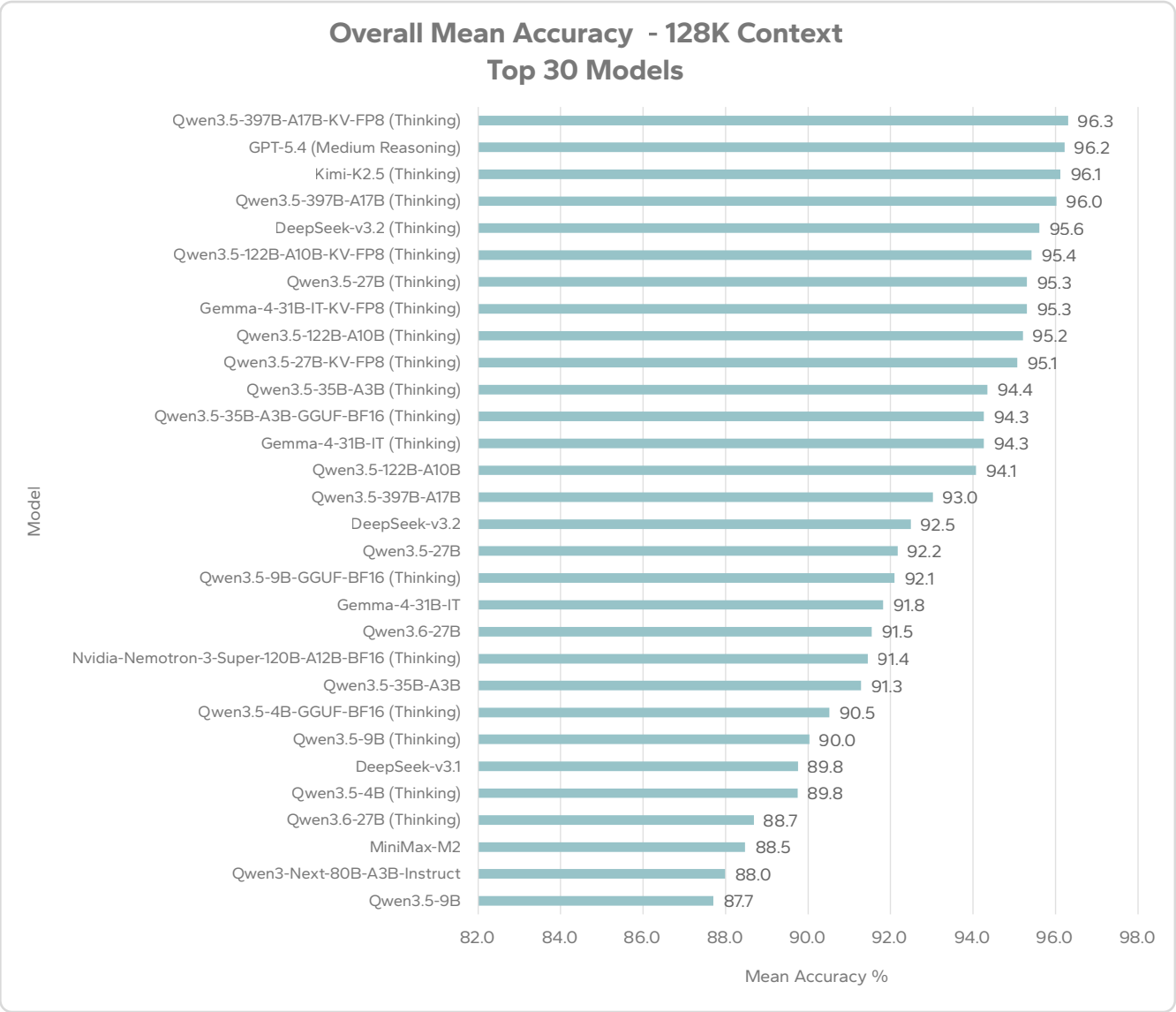


Figure 6: *Overall Mean Accuracy – 128K Context Size*

At 128K context, accuracy declined and model performance diverged. The top 10 models maintained at least 95% accuracy, led by Qwen3.5-397B-A17B-KV-FP8 (Thinking), GPT-5.4 (Medium Reasoning), and Kimi-K2.5 (Thinking). In total, 24 of 72 models exceeded 90%, compared with 45 of 91 at 32K.

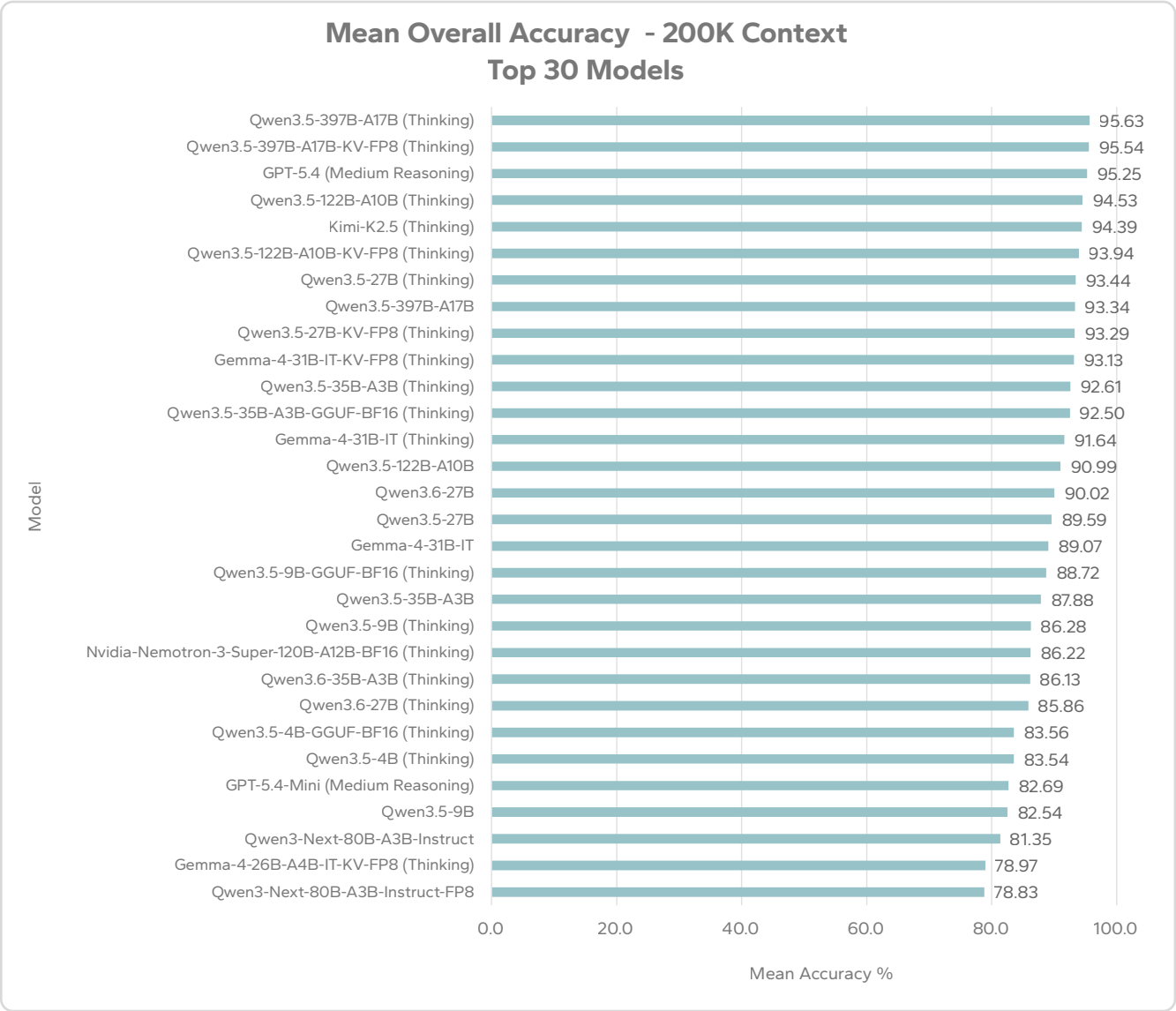


Figure 7: *Overall Mean Accuracy – 200K Context Size*

At 200K context, only three models exceeded 95% overall accuracy: Qwen3.5-397B-A17B (Thinking), Qwen3.5-397B-A17B-KV-FP8 (Thinking), and GPT-5.4 (Medium Reasoning). Twelve more exceeded 90%. The results show that retrieval stability at this scale was limited to a small subset of the 54 models tested.

Context Size	Models > 90% Accuracy	Models > 95% Accuracy
32K	45	27
128K	24	10
200K	15	3

Figure 8: *Context Size Overview Comparison*

Reasoning ("Thinking") vs. Non-Thinking Modes

Thinking mode enables additional model reasoning before the final answer; non-thinking mode answers without that enabled reasoning phase. The comparisons below measure accuracy, not speed or infrastructure efficiency. Teams should evaluate latency, throughput, and cost alongside accuracy when choosing a deployment setting.

Several models were tested in both thinking and non-thinking modes. At 32K context, reasoning modes improved accuracy by 7.13% on average, although the effect varied and a few non-thinking variants performed slightly better.

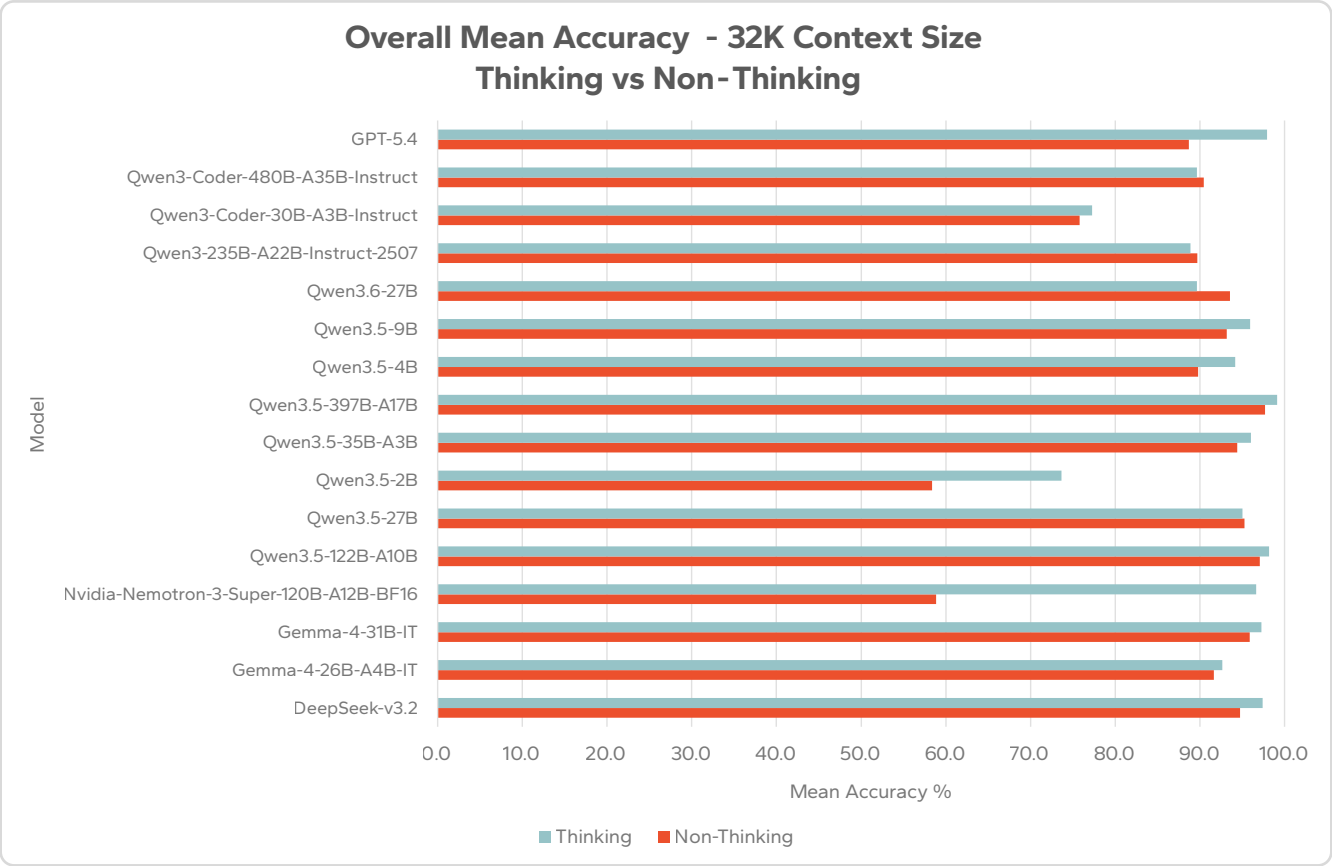


Figure 9: *Thinking vs Non-Thinking Accuracy – 32K Context*

For most models, the change was modest. Qwen3.6-27B (Thinking), for example, scored 4.18% lower than its non-thinking counterpart, while other models improved substantially when reasoning was enabled.

Model	Overall Mean Accuracy - Non-Thinking	Overall Mean Accuracy - Thinking	Percentage Point Gain	Relative Increase
Nvidia-Nemotron-3-Super-120B-A12B-BF16	58.8%	96.6%	+37.8	+64.26%
Qwen3.5-2B	58.4%	73.7%	+15.3	+26.21%
GPT-5.4	88.7%	97.9%	+9.2	+10.43%

Figure 10: *Thinking vs Non-Thinking Comparison – 32K Context*

A similar pattern emerged at longer contexts. Nvidia-Nemotron-3-Super-120B-A12B-BF16 showed the largest relative increase, 89.72%, while many other models improved by 3% to 5%. Because the effect varied by model and context size, organizations should benchmark reasoning modes as distinct deployment configurations.

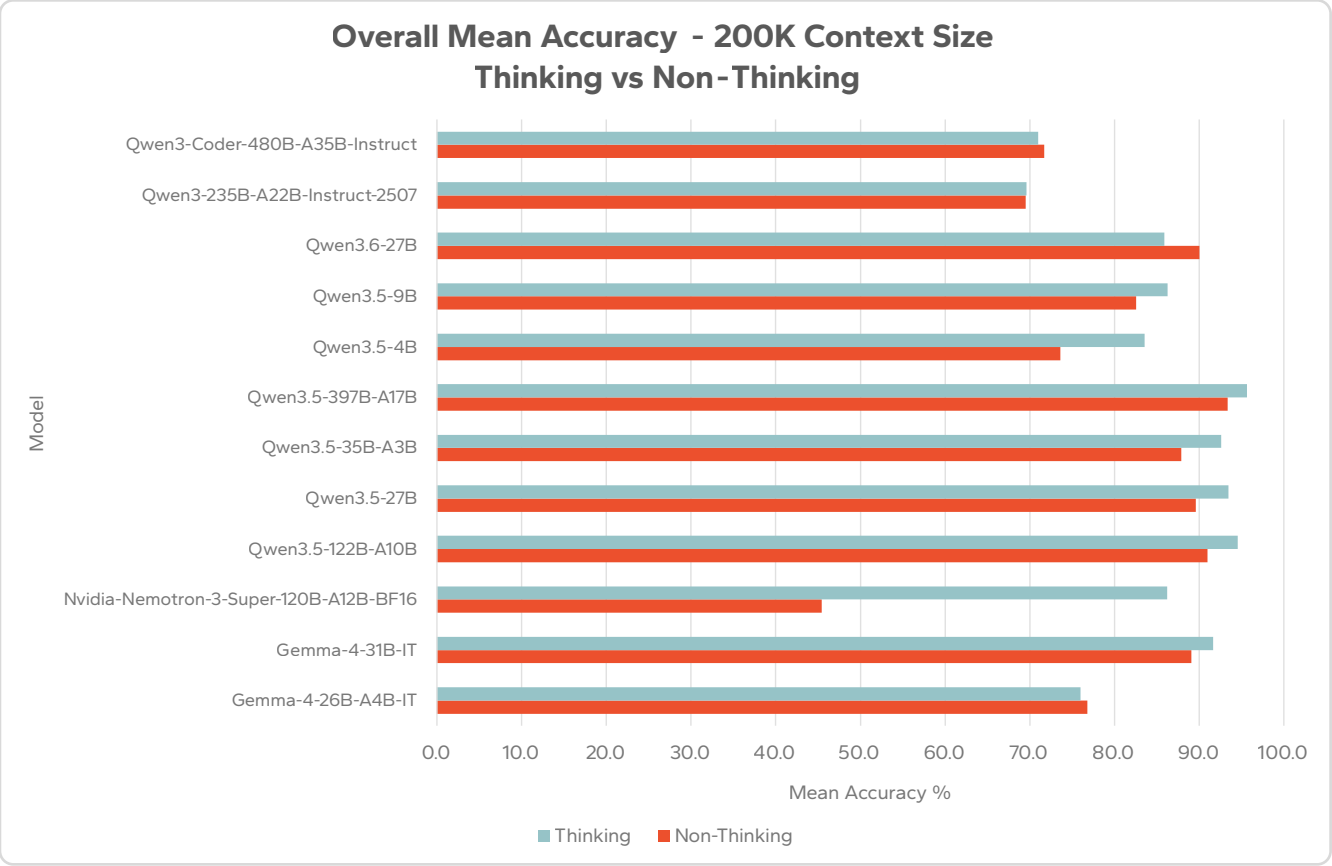


Figure 11: *Thinking vs Non-Thinking Accuracy – 200K Context*

Single-Document Retrieval Tasks

Single-document tasks required models to retrieve and interpret information within one source. At 32K context, 46 models exceeded 90% accuracy and only two scored below 80%.

Accuracy nevertheless declined as context grew. Figure 12 compares the top 20 models across 32K, 128K, and 200K contexts.

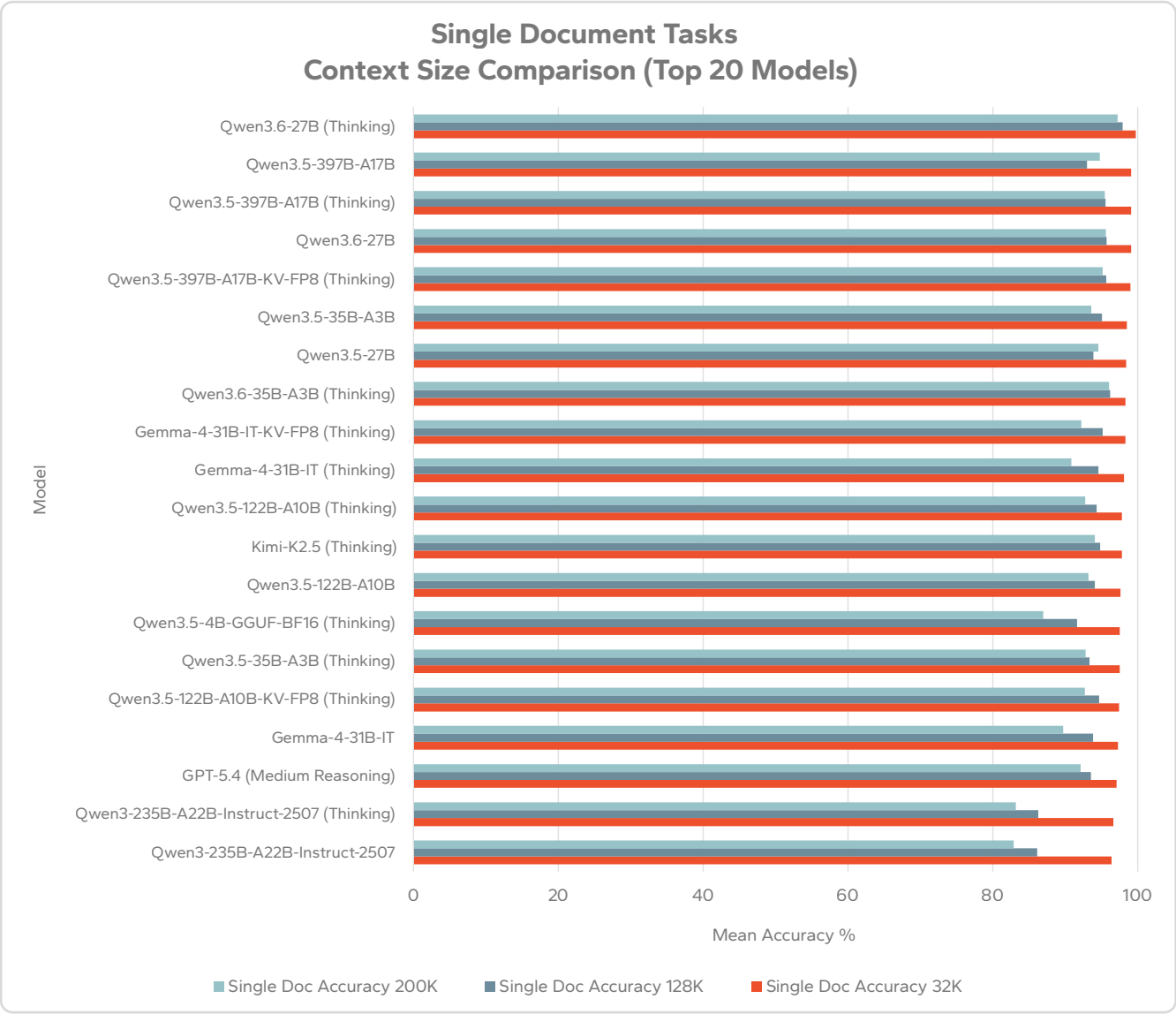


Figure 12: *Single Document Tasks Mean Accuracy*

Among the highest-performing models, retrieval stability remained comparatively strong even at extended context lengths. Qwen3.6-27B (Thinking), Qwen3.5-397B-A17B (Thinking), Qwen3.6-27B, and Qwen3.5-397B-A17B-KV-FP8 (Thinking) all maintained greater than 95% accuracy across all evaluated context sizes, with accuracies declining by only 1.8 to 3.8 percentage points. The non-thinking variation of Qwen3.5-397B-A17B showed the highest degradation within the top five models, decreasing from 99.1% accuracy to 93% and 94.7% accuracies at the two longer context sizes.

Model Name	Single Doc Accuracy 32K	Single Doc Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Single Doc Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Qwen3.6-27B (Thinking)	99.7%	97.9%	-1.8	97.2%	-2.5
Qwen3.5-397B-A17B	99.1%	93.0%	-6.1	94.7%	-4.4
Qwen3.5-397B-A17B (Thinking)	99.1%	95.6%	-3.6	95.4%	-3.7
Qwen3.6-27B	99.1%	95.7%	-3.4	95.6%	-3.5
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	99.0%	95.6%	-3.4	95.2%	-3.8

Figure 13: *Single Document Tasks – Top 5 Models*

Other models degraded sharply. Several systems lost roughly 20 percentage points or more at 200K context, including models that scored 90% to 95% at 32K. GLM-4.6 fell from 94.3% to 53.0%, a 41.3-percentage-point decline, the largest in this category.

Model Name	Single Doc Accuracy 32K	Single Doc Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Single Doc Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Llama-4-Maverick-17B-128E-Instruct	95.7%	80.7%	-15.0	76.0%	-19.7
Gemma-4-26B-A4B-IT (Thinking)	95.6%	84.9%	-10.7	75.2%	-20.5
GLM-4.7	95.3%	87.3%	-8.0	74.9%	-20.4
GLM-4.6	94.3%	89.8%	-4.6	53.0%	-41.3
Gemma-4-26B-A4B-IT	92.4%	82.4%	-10.0	73.7%	-18.7
Qwen3-4B-Instruct-2507	89.0%	73.6%	-15.4	57.9%	-31.1
Llama-4-Scout-17B-16E-Instruct	84.1%	71.9%	-12.2	58.4%	-25.8
Qwen3.5-2B (Thinking)	74.5%	61.3%	-13.2	54.8%	-19.7
Qwen3.5-0.8B	71.5%	56.4%	-15.1	53.8%	-17.7

Figure 14: *Single Document Tasks – Decrease Across Context Sizes*

Multi-Document Aggregation Tasks

Multi-document aggregation was substantially harder than single-document extraction. These tasks required models to compare, summarize, and derive relationships across sources, a common pattern in enterprise retrieval workflows.

Aggregation accuracy degraded much faster as context grew. At 32K, the top 20 models all exceeded 95%; at 128K, only five remained above 90%. At 200K, no model surpassed 95%. Figure 15 compares aggregation performance across context sizes.

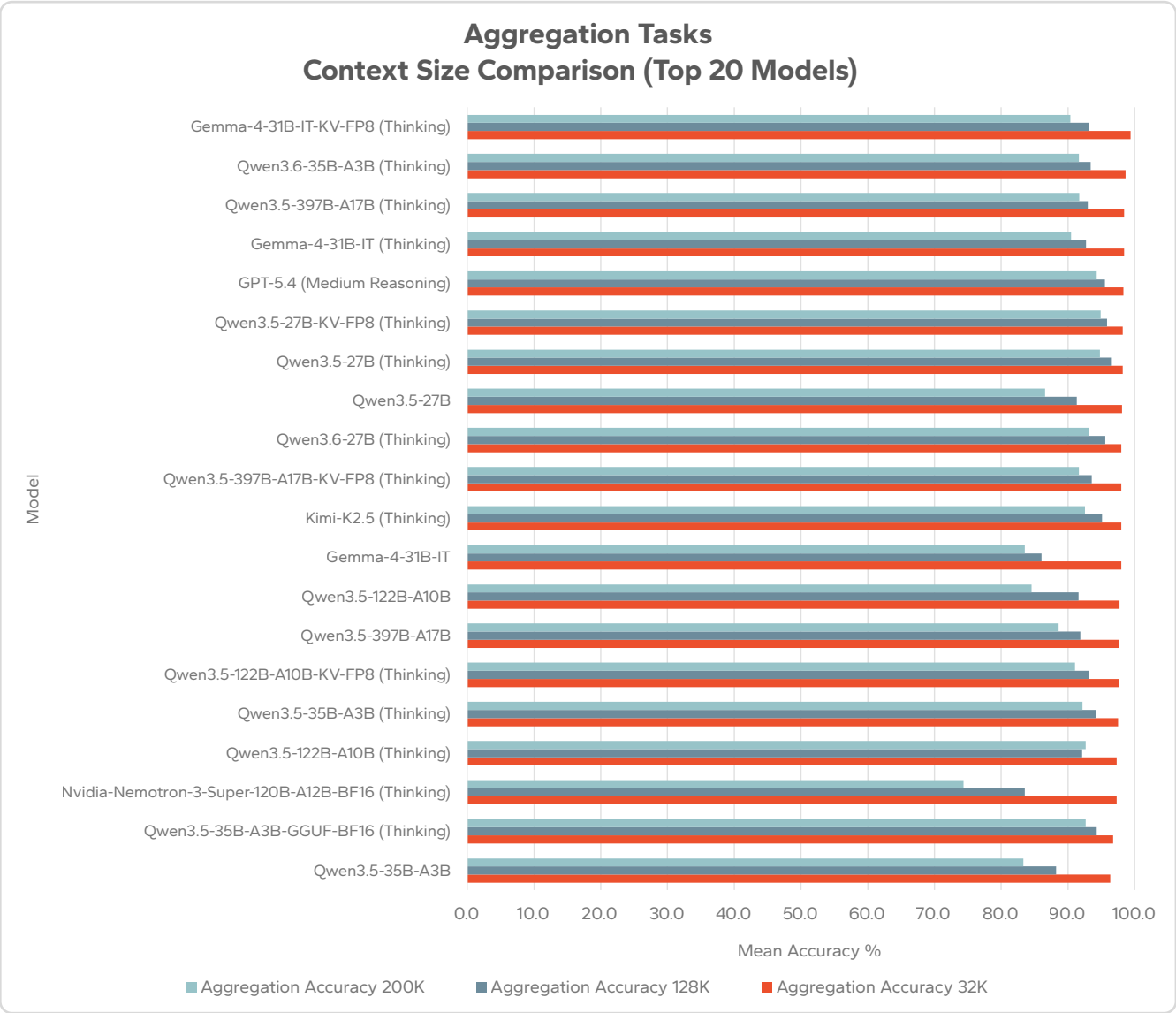


Figure 15: *Multi-document Aggregation Tasks*

Among the five highest-performing models shown, accuracy declined by 2.81 to 6.32 percentage points at 128K and by 4.0 to 9.0 points at 200K, relative to 32K.

Model name	Aggregation Accuracy 32K	Aggregation Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Aggregation Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Gemma-4-31B-IT-KV-FP8 (Thinking)	99.4%	93.0%	-6.32	90.3%	-9.0
Qwen3.6-35B-A3B (Thinking)	98.6%	93.4%	-5.24	91.6%	-7.0
Gemma-4-31B-IT (Thinking)	98.4%	92.7%	-5.73	90.5%	-8.0
Qwen3.5-397B-A17B (Thinking)	98.4%	93.0%	-5.45	91.7%	-6.7
GPT-5.4 (Medium Reasoning)	98.3%	95.5%	-2.81	94.3%	-4.0

Figure 16: *Multi-document Aggregation Tasks – Top 5 Models*

Across all evaluated models, average aggregation accuracy declined by 16.5% at 128K and 25.6% at 200K relative to 32K performance. That average degradation was more than twice the rate observed for single-document tasks, making cross-document aggregation the benchmark's clearest production-readiness bottleneck.

Model name	Aggregation Accuracy 32K	Aggregation Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Aggregation Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Nvidia-Nemotron-3-Super-120B-A12B-BF16 (Thinking)	97.3%	83.5%	-13.8	74.3%	-22.9
Gemma-4-26B-A4B-IT-KV-FP8 (Thinking)	95.6%	85.5%	-10.1	76.0%	-19.6
Qwen3-Next-80B-A3B-Instruct	93.9%	81.3%	-12.6	64.0%	-29.9
Qwen3-235B-A22B-Instruct-2507	93.7%	76.4%	-17.3	58.8%	-34.9

Model Name	Aggregation Accuracy 32K	Aggregation Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Aggregation Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Qwen3-Coder-480B-A35B-Instruct	93.4%	73.5%	-19.9	60.7%	-32.7
Qwen3.5-4B	92.1%	76.6%	-15.5	64.1%	-28.0
GLM-4.6	91.8%	78.3%	-13.5	23.7%	-68.1
Qwen3-Coder-480B-A35B-Instruct (Thinking)	91.6%	72.8%	-18.8	61.5%	-30.1
Llama-4-Maverick-17B-128E-Instruct	91.0%	49.4%	-41.6	51.2%	-39.7
GLM-4.7	81.2%	74.9%	-6.3	42.8%	-38.3

Figure 17: *Multi-document Aggregation Tasks – Top 10 Models*

Llama-4-Maverick-17B-128E-Instruct fell from 91.0% at 32K to 49.4% at 128K. GLM-4.6 showed the largest overall reduction, declining by 68.1 percentage points between 32K and 200K.

Cross-document reasoning and synthesis therefore remain major challenges for long-context systems, including models that perform strongly on simpler retrieval tasks.

Hallucination-Probing Tasks

The final category probes whether models generate information absent from the corpus. One task asks about nonexistent entities; the other asks about optional fields intentionally omitted from specific documents. Correct responses acknowledge the absence with answers such as “Unknown” or “N/A.”

Hallucination-probe accuracy was more stable than retrieval and aggregation accuracy as context grew. Qwen3.5-397B-A17B (Thinking) exceeded 99% across all evaluated context sizes, and several other top models remained similarly stable.

Some models still degraded substantially. GLM-4.6, for example, fell by 61.2 percentage points between 32K and 200K.

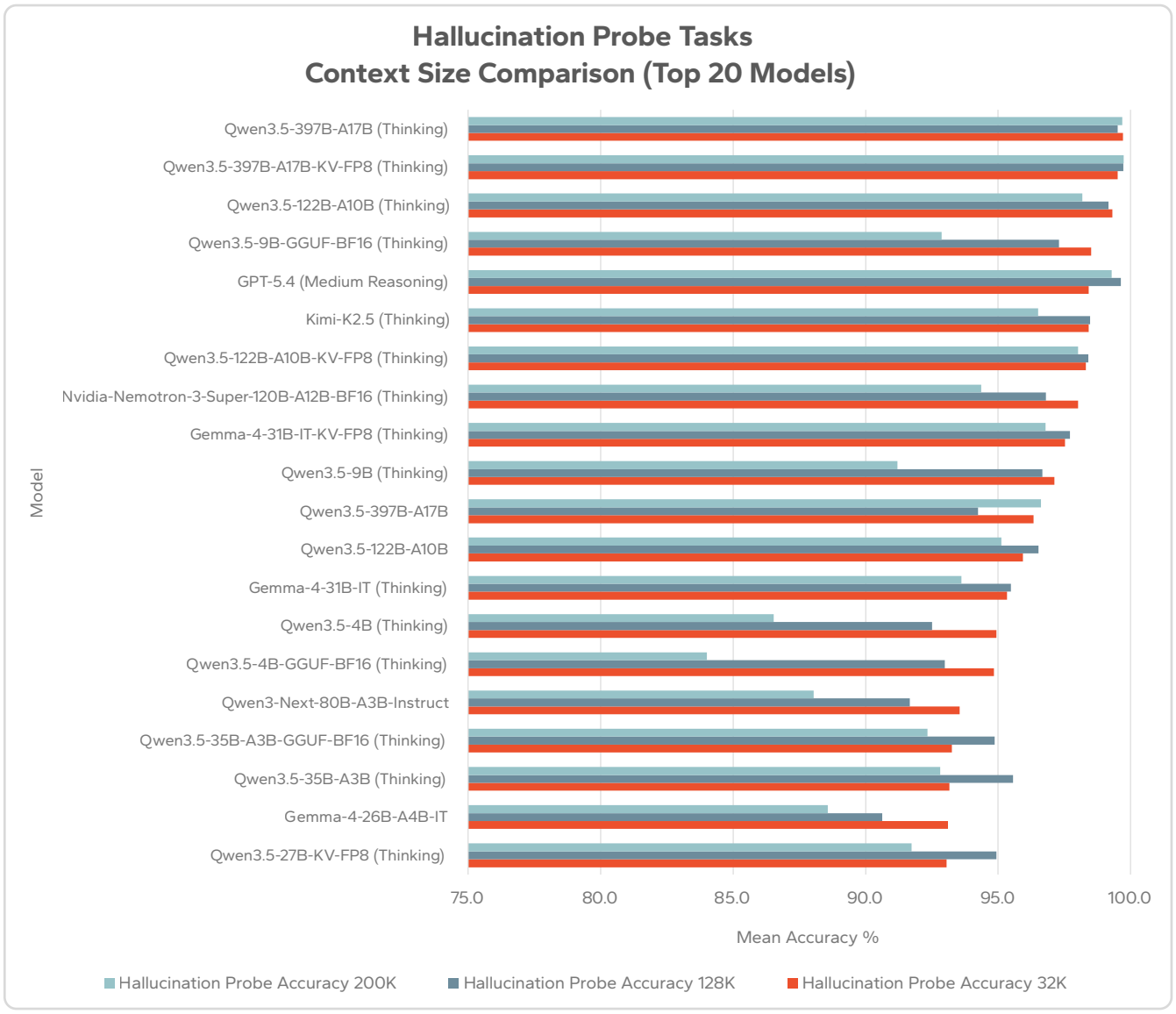


Figure 18: *Hallucination Probing Tasks*

Model Name	Hallucination Probe Accuracy 32K	Hallucination Probe Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Hallucination Probe Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Qwen3.5-397B-A17B (Thinking)	99.7%	99.5%	-0.2	99.7%	0.0
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	99.5%	99.7%	+0.2	99.7%	+0.2
Qwen3.5-122B-A10B (Thinking)	99.3%	99.2%	-0.1	98.2%	-1.1

Model Name	Hallucination Probe Accuracy 32K	Hallucination Probe Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Hallucination Probe Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Qwen3.5-9B-GGUF-BF16 (Thinking)	98.5%	97.3%	-1.2	92.9%	-5.6
GPT-5.4 (Medium Reasoning)	98.4%	99.6%	+1.2	99.3%	+0.9
Kimi-K2.5 (Thinking)	98.4%	98.5%	+0.1	96.5%	-1.9
GLM-4.6	91.7%	84.8%	-6.9	30.5%	-61.2

Figure 19: *Hallucination Probing Tasks - Context Size Comparison*

Relative to 32K, average hallucination-probe accuracy declined by 2.79% at 128K and 9.78% at 200K, less than in either retrieval category.

The results suggest that "can the model find the answer?" and "does it recognize when the answer is absent?" may be partially independent capabilities. Many models became worse at retrieval and aggregation as context grew without becoming equivalently more likely to invent missing information.

Complete per-model results across all context sizes and task categories are available from Signal65.

Operational Implications for Enterprise AI

Use RIKER to shortlist models and reasoning settings at the document input sizes your application will supply, then validate those choices in the complete retrieval or agent workflow. The following seven actions guide that process.

Benchmark at application input sizes – Treat advertised context windows as capacity limits, not evidence of retrieval quality. Test the amount and mix of document text supplied to the model in the intended workflow.

Test aggregation separately from lookup – Include totals, comparisons, enumeration, and temporal questions across documents. Simple single-document pilots can materially overstate production readiness.

Prioritize stability over peak short-context scores – Compare accuracy curves across workload sizes; stable performance at the target scale is more operationally useful than a marginal lead at 32K.

Test reasoning modes independently – Treat thinking and non-thinking variants as separate configurations. Measure quality together with latency, throughput, and cost before selecting a production setting.

Measure abstention separately – Evaluate both answer retrieval and the model's ability to recognize absent information; one score should not substitute for the other.

Validate retrieval before adding autonomy – Test production-scale retrieval and aggregation before selecting agent frameworks or delegating consequential actions.

Validate the complete deployment stack – Long-context inference can be memory- and communication-intensive. Size compute, host and GPU memory, and the network fabric against the models, context sizes, concurrency, latency, and throughput targets the organization intends to run. The Dell PowerEdge XE9680 and XE7745 with Broadcom 400GbE RoCEv2 fabric formed the tested open-model stack; this benchmark did not isolate infrastructure components or establish that this specific configuration is required.

What This Means for the Industry

Beyond individual model performance, this testing also highlights broader issues in how the AI industry evaluates and markets long-context capability. Three patterns deserve attention beyond any individual model score.

The context-window arms race has outpaced enterprise usability. Advertised token capacity does not establish retrieval quality. Of 54 models tested, only three exceeded 95% overall accuracy at 200K, while average aggregation accuracy declined by 25.6% relative to 32K. Vendors and evaluators should publish retrieval-stability curves across context scale, not token limits alone.

Enterprise evaluation needs production-relevant tasks. Public leaderboards can be limited by training-data contamination, judge-model variability, and tasks that poorly approximate enterprise retrieval. RIKER addresses these issues with inverted generation, deterministic grading, and coherent corpora modeled on enterprise document types. The industry should evaluate retrieval, aggregation, and abstention as distinct capabilities.

Enterprise AI requires coordinated benchmarks. RIKER measures retrieval, aggregation, and responses to absent information. KAMI evaluates whether models can plan, act, and recover in enterprise workflows. PINNACLE is intended to extend the framework to additional performance and quality dimensions. No single benchmark can establish production readiness; results should be combined with workload-specific system testing.

Limitations and Future Work

While the benchmark is designed to approximate realistic enterprise retrieval workloads, several limitations remain. Although the benchmark utilizes coherent synthetic document corpora to reduce inconsistencies commonly found in synthetic datasets, production enterprise environments may contain substantially greater variability, noise, and domain-specific complexity.

The current benchmark corpus also focuses on a limited set of enterprise domains: HR records, commercial leases, and facility field reports. Performance may differ in specialized domains such as healthcare, legal analysis, scientific research, or software engineering.

Current evaluations are also limited to retrieval over provided context windows rather than complete end-to-end enterprise retrieval architectures. Real-world deployments frequently incorporate additional systems such as retrieval-augmented generation (RAG) pipelines, vector databases, knowledge graphs, and agentic retrieval workflows, each of which may introduce additional retrieval and reasoning challenges.

Future work will expand RIKER and KAMI to additional enterprise workloads, retrieval architectures, and long-context reasoning scenarios. Signal65 and Kamiwaza are also developing PINNACLE to evaluate additional dimensions of model quality and deployment performance across common enterprise use cases and personas.

Conclusion

Enterprise AI knowledge systems promise faster access to institutional knowledge, but that promise should be evaluated rather than assumed. Models that appeared similar at smaller contexts diverged as workloads grew; average aggregation accuracy declined by 25.6% at 200K relative to 32K.

For leaders evaluating AI investments, the guidance is practical: test models at the organization's actual document scale; evaluate cross-document aggregation directly; treat reasoning settings as distinct deployment configurations; and measure whether the system recognizes absent information. These factors are more decision-useful than advertised context limits alone.

The implications are especially important for agentic AI: agents act on retrieved information, so retrieval errors can become action errors. RIKER evaluates the foundation, trusted answers, while Signal65 and Kamiwaza's KAMI benchmark evaluates the next layer: planning, acting, and recovering within enterprise workflows.

A small set of models demonstrated strong retrieval and abstention performance across every tested workload size. The gap between those leaders and the broader field makes model and configuration selection a decision with measurable operational consequences. RIKER provides a repeatable framework for comparing those choices under controlled conditions.

The results also reinforce the importance of evaluating AI infrastructure as a system. The tested open-model environment paired Dell PowerEdge XE9680 and XE7745 compute and memory capacity with Broadcom BCM57608-based 400GbE RoCEv2 connectivity through a Dell Z9864F-ON switch. This configuration supported the evaluated workloads in the Signal65 lab; production sizing should be validated against each organization's models, context sizes, concurrency, latency, throughput, security, and cost requirements.

Reliable retrieval is becoming a prerequisite for trusted enterprise AI. Organizations should evaluate models, configurations, and infrastructure together against the workloads they actually intend to run—not against specifications alone.

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