signal65

Benchmarking Information Retrieval and LLM Hallucination with RIKER

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

IN PARTNERSHIP WITH

KAMI WAZA

JUNE 2026

Executive Summary

Information retrieval and contextual understanding are foundational components of enterprise AI workflows. By incorporating additional context through approaches such as retrieval-augmented generation (RAG), AI models can provide domain-specific and document-specific assistance. For enterprises deploying these systems, however, understanding how effectively models retrieve information, and when they are prone to hallucination, is increasingly critical.

As part of an ongoing collaboration between Signal65 and Kamiwaza, this paper introduces RIKER (Retrieval Intelligence and Knowledge Extraction Rating), a benchmark designed to evaluate LLM knowledge retrieval capabilities. RIKER builds upon and complements the KAMI benchmark, a previous Signal65 and Kamiwaza effort focused on evaluating agentic AI capability. Together, the RIKER and KAMI benchmarks provide insight into how LLMs perform across common enterprise AI workloads.

Key findings of the RIKER benchmark include:

- **Qwen3.5-397B-A17B (Thinking) achieved the strongest overall retrieval performance**: This model achieved the highest overall retrieval performance of all models tested, and was among a handful of models that consistently maintained high accuracy across retrieval tasks and context sizes, while many other models experienced substantial degradation as context length increased. Other top models include Gemma-4-31B-IT-KV-FP8 (Thinking), Qwen3.5-122B-A10B (Thinking), Kimi-K2.5 (Thinking), and GPT-5.4 (Medium Reasoning), all of which maintained greater than 94% overall accuracy at 200K context.
- **Long context windows are not equivalent to reliable retrieval:** Although modern LLMs support increasingly large context sizes, retrieval accuracy consistently declined as context length expanded.
- **Aggregation tasks degrade faster than single-document retrieval:** Models experienced significantly greater accuracy loss when required to aggregate or compare information across multiple documents versus retrieving information from a single source.
- **Effective long-context retrieval varies dramatically across models:** At a 32K context size, 27 models achieved overall accuracy above 95%. At a 200K context size, that number fell to just 3 models. While many models perform similarly at moderate context lengths, performance divergence increases substantially as context size grows.

Key Takeaways

- **91 models benchmarked** across real-world enterprise retrieval workloads
- 27 models **exceeded 95% accuracy** at 32K context
- Only 3 models remained **above 95% accuracy** at 200K context
- Multi-document aggregation accuracy **declined more than 2× faster** than single-document retrieval
- Thinking models improved retrieval performance by **up to 64%**

- **Thinking can improve information retrieval:** The top performing models across all context lengths were comprised of thinking models. When compared to non-thinking variations of the same models, thinking models typically demonstrated improved performance.
- **Hallucination behavior appears less sensitive to context length than retrieval accuracy:** While information retrieval and aggregation performance declined substantially as context size increased, hallucination-probing tasks exhibited comparatively smaller changes. This suggests that retrieval failures and hallucination behavior may represent partially distinct failure modes in long-context LLM systems.

What does this mean for the AI analysis and performance testing industry?

- **The context window arms race is outpacing enterprise usability:** vendors compete on token caps, but only three of 54 models tested sustained production-grade accuracy at 200K. Advertised context length is not a proxy for usable retrieval.
- **Public benchmarks are part of the problem:** training contamination, LLM-as-judge subjectivity, and synthetic extraction tasks have produced leaderboards that don't predict enterprise outcomes. RIKER's inverted-generation and deterministic grading are a direct response.
- **RIKER is one pillar of a coordinated evaluation framework:** paired with KAMI for agentic capability and our upcoming project PINNACLE, Signal65 and Kamiwaza are building the enterprise-grade measurement stack the field has lacked.

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. Information can be provided to LLMs in many different ways, including simply providing relevant documents during a chat session, to more complex retrieval systems such as RAG, knowledge graphs, or agentic tooling. Information retrieval forms the backbone of most enterprise AI workloads that are being deployed today; however, it also introduces new challenges related to retrieval accuracy, reasoning across documents, 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, KAMI and RIKER 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, RIKER evaluates a model's ability to retrieve, aggregate, and reason over information distributed across realistic document corpora.

RIKER differentiates itself from many existing benchmarks 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, RIKER 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.

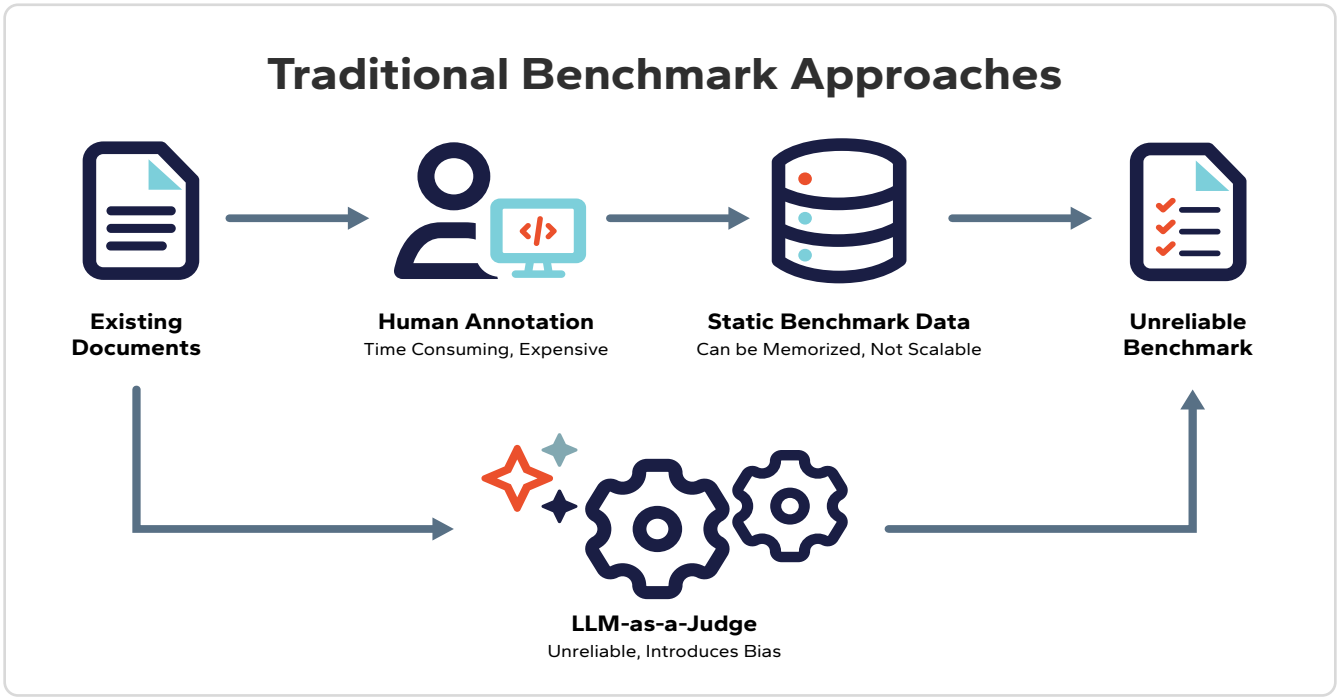


Figure 1: *Traditional Benchmark Approaches*

RIKER 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 RIKER to generate unique benchmark instances that are resistant to memorization while maintaining scalable, repeatable scoring.

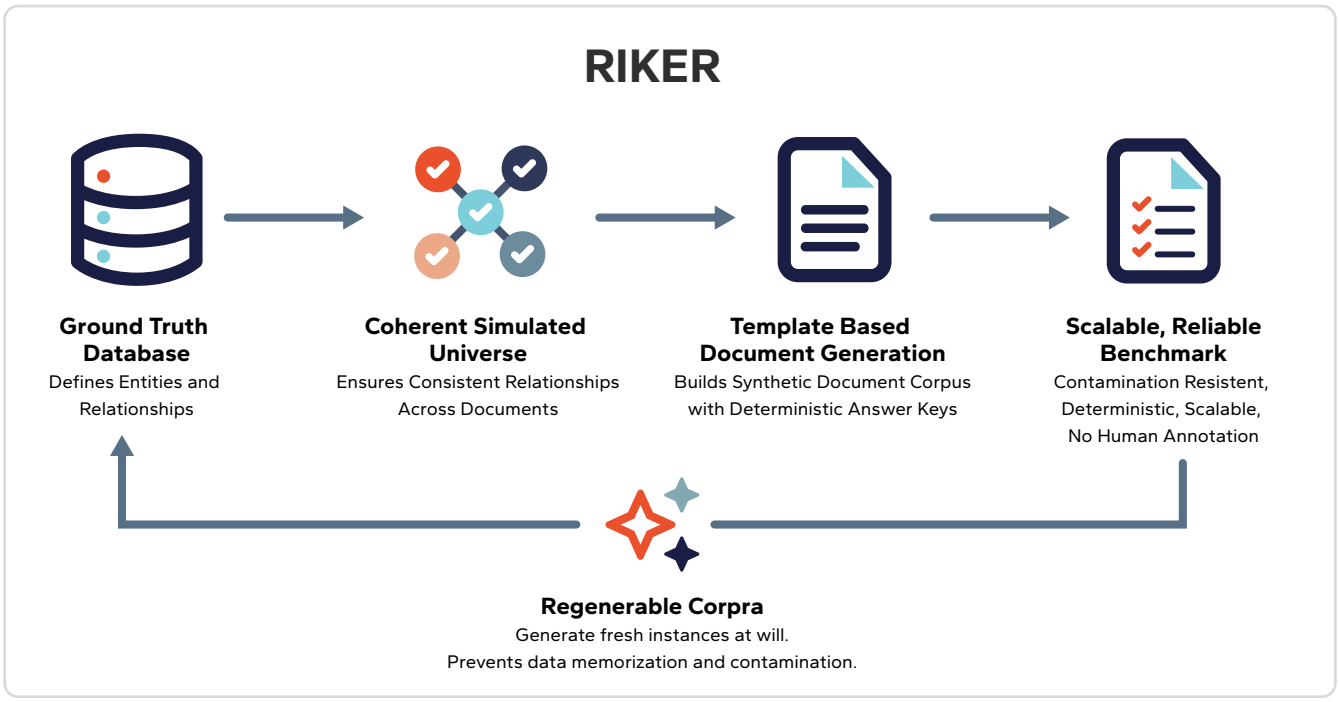


Figure 2: *RIKER Overview*

A key feature of RIKER is its "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. RIKER instead maintains consistent relationships between entities across documents, enabling coherent multi-document retrieval and aggregation tasks that better approximate enterprise knowledge environments.

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

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 workloads. The benchmark additionally includes hallucination probing tasks designed to evaluate whether models incorrectly generate information that does not exist within the provided corpus.

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

Section	Question #	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 Lessor 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 Lessor 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. To evaluate the impact of context length on retrieval performance, models were tested with corpora at three distinct context sizes: 32K, 128K, and 200K tokens. In total, testing included 91 models at 32K context, 72 models at 128K context, and 54 models at 200K context.

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.

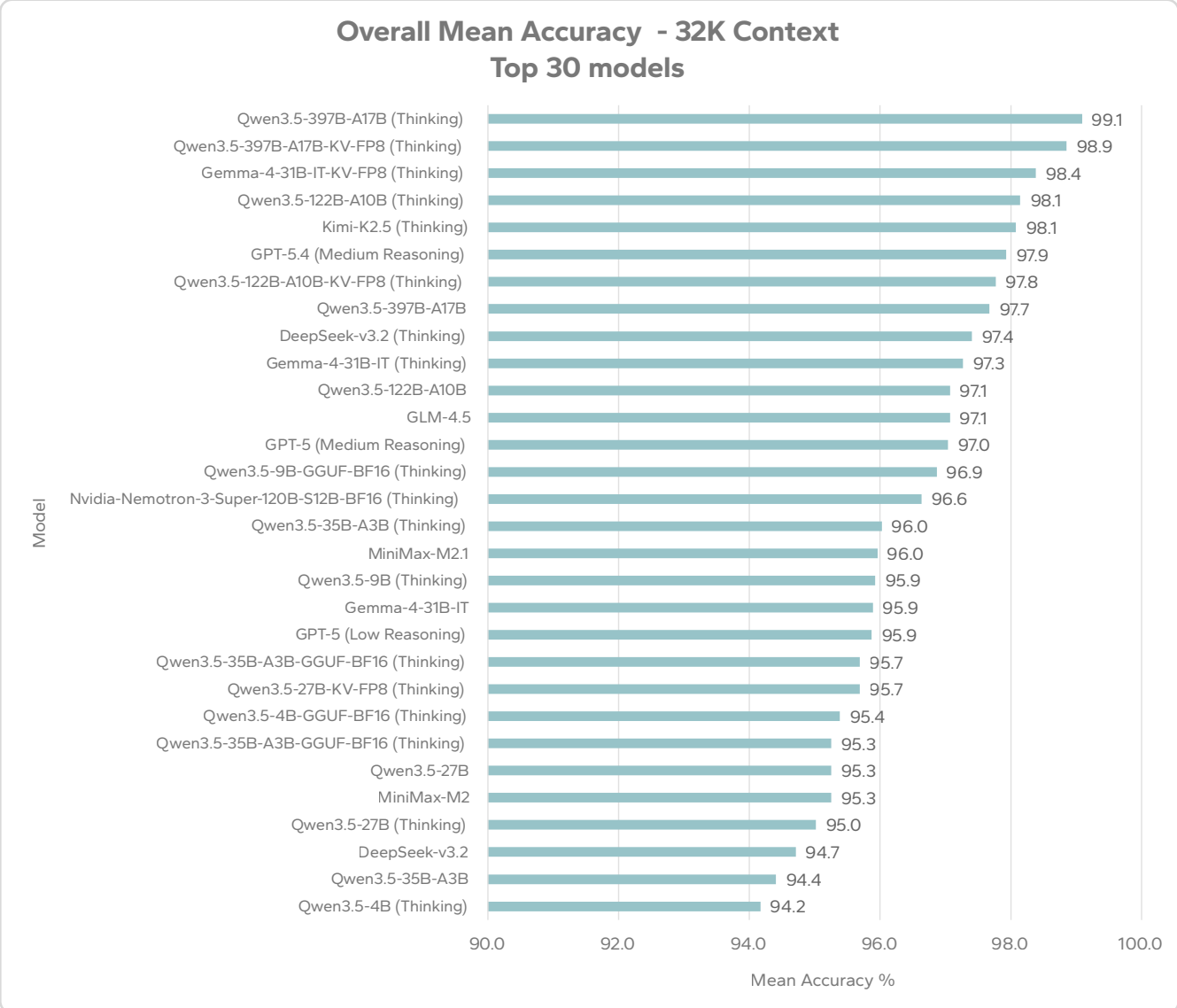


Figure 4: *Overall Mean Accuracy – 32K Context Size*

At a 32K context size, retrieval performance remained relatively strong across a wide range of models. All models within the top 30 achieved accuracy greater than 90%, with 27 achieving 95% or more and five surpassing 98%. Qwen3.5-397B-A17B (Thinking) achieved the highest overall score at 99.1%, closely followed by its FP8 quantized variation at 98.9%. Several additional models – including Gemma-4-31B-IT-KV-FP8 (Thinking), Qwen3.5-122B-A10B (Thinking), Kimi-K2.5 (Thinking), and GPT-5.4 (Medium Reasoning) – achieved similarly strong performance. Overall, score distributions at 32K context remained relatively compressed, with 45 of the 91 models tested achieving accuracy of 90% or higher. This indicates that a wide range of models can effectively perform moderate-scale retrieval workloads.

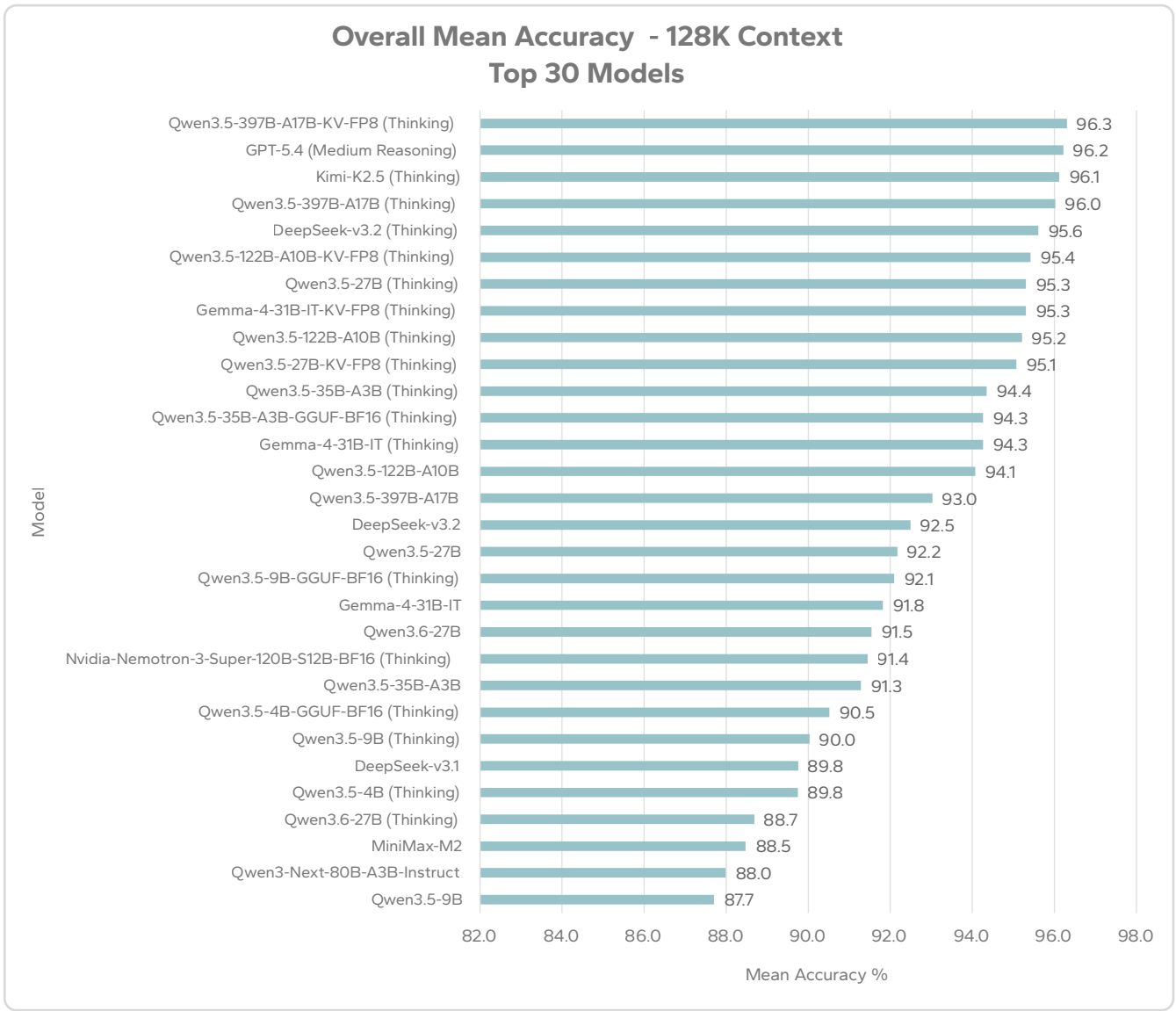


Figure 5: *Overall Mean Accuracy – 128K Context Size*

At 128K context, model accuracies began to degrade and performance divergence increased. The top 10 models all maintained accuracy of 95% or more, led by Qwen3.5-397B-A17B (Thinking), GPT-5.4 (Medium Reasoning), and Kimi-K2.5 (Thinking). In total, 24 models achieved accuracy above 90%, however, this represents only 33% of models tested. This marks a notable decline compared to 32K context testing, in which nearly 50% of models tested surpassed the 90% accuracy threshold.

This degradation can be noted within the top 30 models, in which model accuracies begin falling below 90%. Many additional models that previously performed strongly at 32K context experienced further degradation, dropping to accuracies of 70% and lower. These results suggest that effective long-context retrieval capability varies significantly between models, even among systems advertising similar context limits.

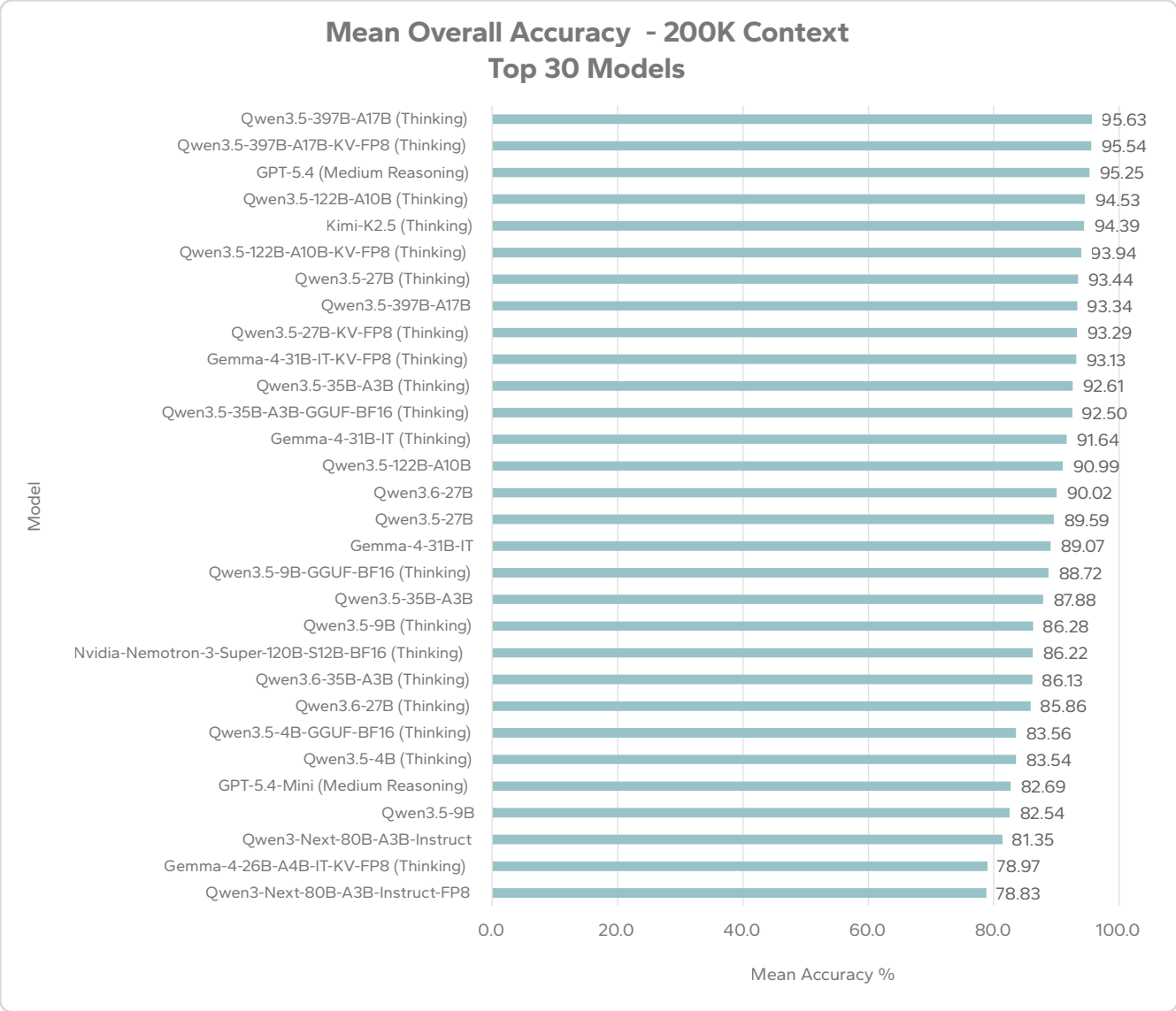


Figure 6: *Overall Mean Accuracy – 200K Context Size*

Accuracy declined further at 200K context, with only three models achieving scores above 95% overall accuracy: Qwen3.5-397B-A17B (Thinking), Qwen3.5-397B-A17B-KV-FP8 (Thinking), and GPT-5.4 (Medium Reasoning). Notably, these models demonstrated relatively limited degradation across all evaluated context sizes. An additional 12 models remained above 90% accuracy. The consistent decline in accuracy across all evaluated context lengths suggests that effective LLM knowledge retrieval remains highly sensitive to context scale, while stable long-context retrieval capability is currently limited to a small subset of models.

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

Figure 7: *Context Size Overview Comparison*

Thinking vs Non-Thinking Models

Several models were tested in both thinking and non-thinking modes. In general, thinking models were found to consistently outperform their non-thinking variations. At a 32K context size, thinking was found to improve accuracy by 7.13% on average.

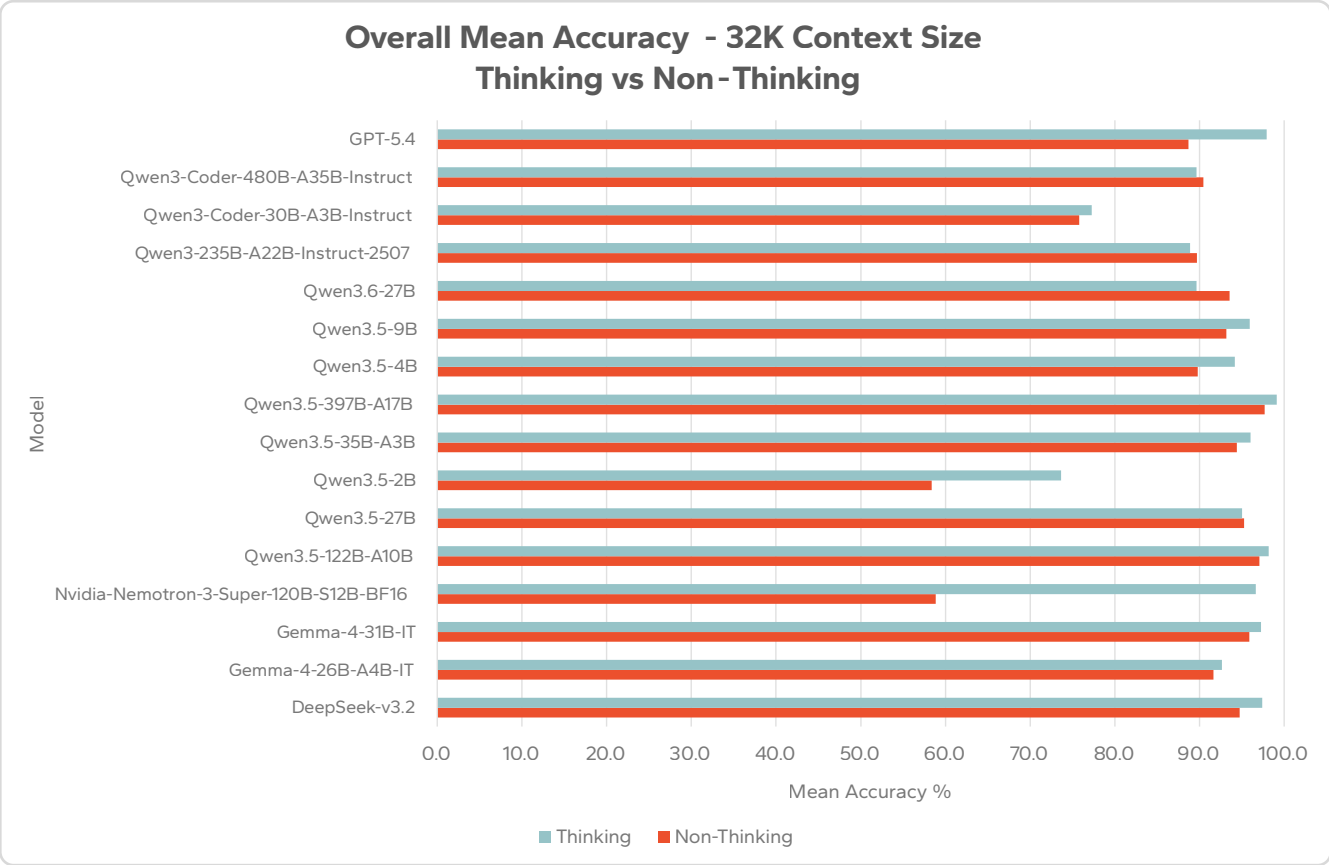


Figure 8: *Thinking vs Non-Thinking Accuracy – 32K Context*

For most models, the change in accuracy was relatively modest, and in a few cases non-thinking models slightly outperformed their thinking counterparts. Most notably, Qwen3.6-27B (Thinking) declined 4.18% compared to its non-thinking mode. Certain models, however, experienced significant accuracy improvements when tested with thinking enabled.

Model	Overall Mean Accuracy - Non-Thinking	Overall Mean Accuracy - Thinking	Percentage Point Gain	Relative Increase
Nvidia-Nemotron-3-Super-120B-S12B-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 9: *Thinking vs Non-Thinking Comparison – 32K Context*

A similar trend was additionally found at longer context lengths. The most notable improvement was once again Nvidia-Nemotron-3-Super-120B-S12B-BF16 which increased its accuracy by 89.72%. For most other models, thinking improved accuracy less dramatically, however many models still experienced improvements between 3% and 5%. For enterprises evaluating models to deploy in production environments, even relatively small improvements can be impactful.

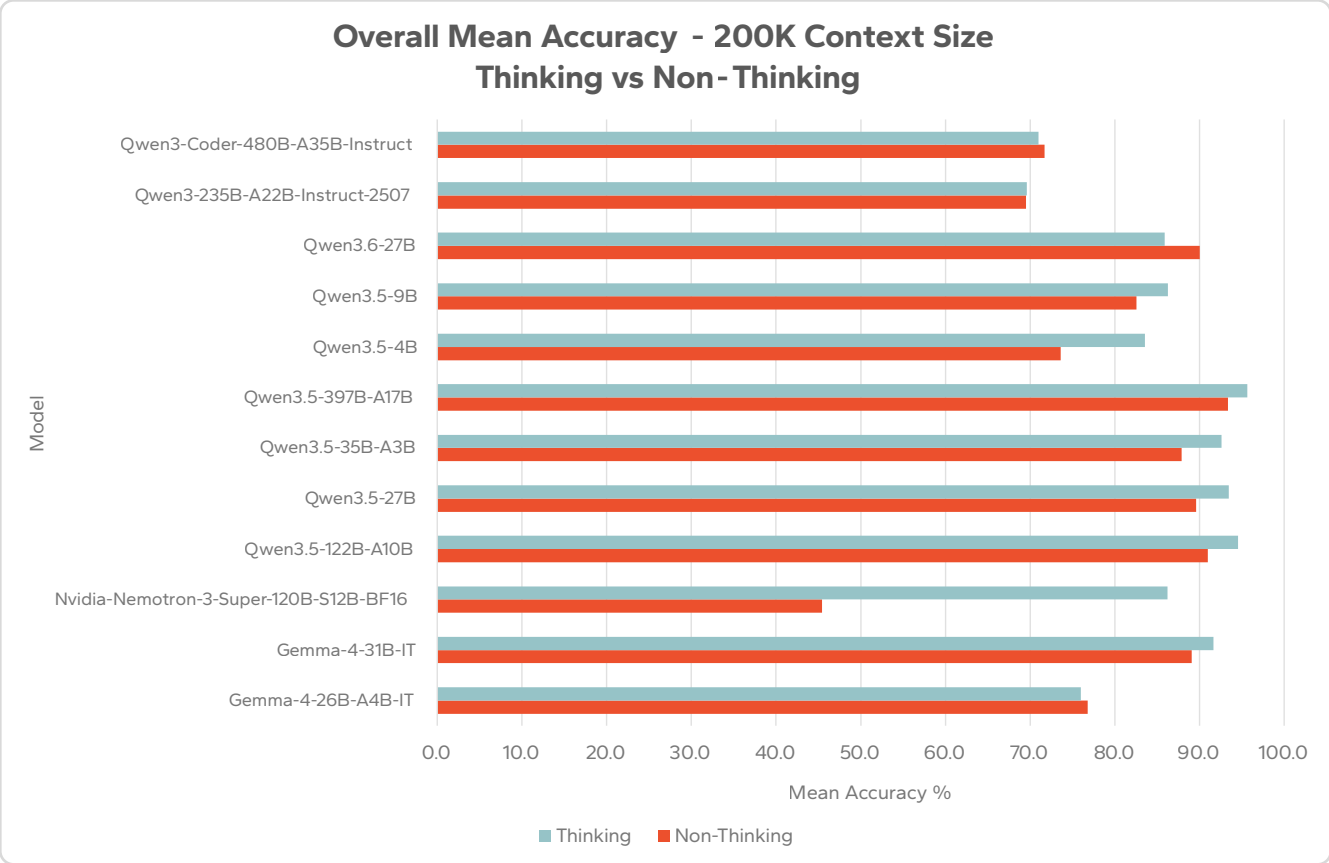


Figure 10: *Thinking vs Non-Thinking Accuracy – 200K Context*

Single Document Retrieval Tasks

Single-document retrieval tasks represented the least complex evaluation category in the RIKER benchmark, requiring models to retrieve and interpret information contained within a single source document. At 32K context, performance across these tasks remained relatively strong for most evaluated systems, with 46 models achieving greater than 90% accuracy and only 2 models scoring below 80%.

Despite the relative simplicity of these workloads, retrieval accuracy consistently declined as context length increased. Figure 11 illustrates single-document retrieval performance across 32K, 128K, and 200K context sizes for the top 20 highest performing models.

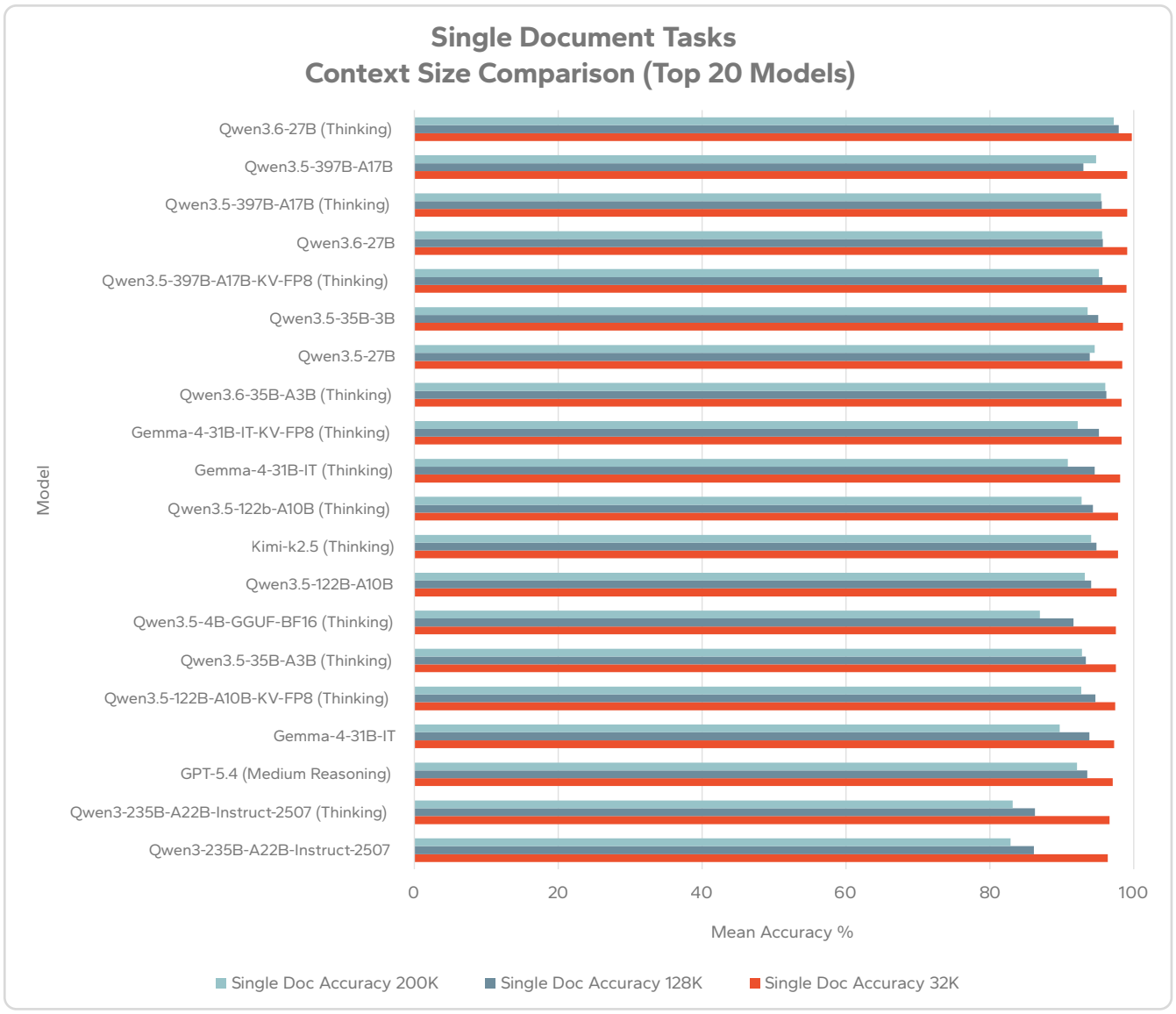


Figure 11: *Single Document Tasks Mean Accuracy*

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, 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 12: *Single Document Tasks – Top 5 Models*

Other models, however, experienced substantially greater degradation as context size increased. Several lower-performing systems — including Qwen3.5-0.8B, Qwen3.5-2B (Thinking), and Llama-4-Scout-17B-16E-Instruct — experienced accuracy declines approaching or exceeding 20 percentage points when evaluated at 200K context. More notably, however, several models that initially achieved 90-95% accuracy at 32K context experienced significant degradation at larger context sizes. GLM-4.6, for example, declined from 94.3% accuracy at 32K context to 53% at 200K, a loss of 41.3 percentage points – the largest observed decrease 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

Model Name	Single Doc Accuracy 32K	Single Doc Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Single Doc Accuracy 200K	Accuracy Change (32K -> 200K, pp)
Qwen3-4B-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 13: *Single Document Tasks – Decrease Across Context Sizes*

Multi-Document Aggregation Tasks

Multi-document aggregation tasks represented a substantially more difficult retrieval workload than single-document extraction. These tasks required models to compare, summarize, and derive relationships across multiple documents within the provided corpus. This more closely approximates most common enterprise retrieval workflows where relevant information is distributed across many sources.

Compared to single-document retrieval tasks, models demonstrated significantly greater degradation on aggregation workloads as context length increased. At 32K context, aggregation performance remained relatively strong, with the top 20 models all achieving greater than 95% accuracy. At 128K context, however, only five models — GPT-5.4 (Medium Reasoning), Qwen3.5-27B-KV-FP8 (Thinking), Qwen3.5-27B (Thinking), Qwen3.6-27B (Thinking), and Kimi-k2.5 (Thinking) — maintained accuracy above 90%. No models surpassed 95% accuracy at the 200K context size. A comparison of aggregation performance across context sizes can be seen in Figure 14.

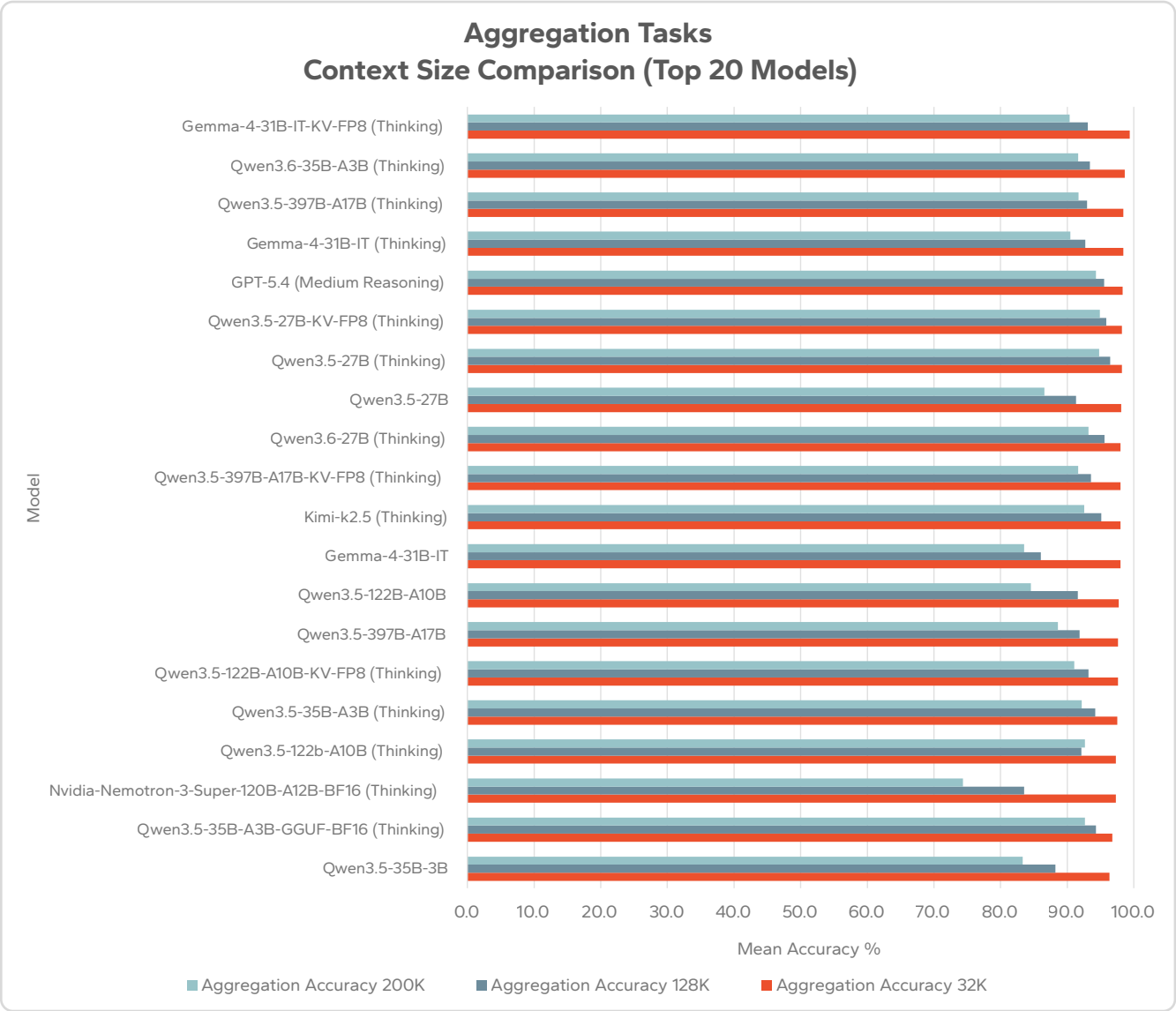


Figure 14: *Multi-document Aggregation Tasks*

Among the highest-performing models, loss of accuracy is moderate, ranging from a loss of 2.81 to 6.32 percentage points when context is extended to 128K, and between 4 and 9 percentage points when extended to 200K.

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 15: *Multi-document Aggregation Tasks – Top 5 Models*

Beyond the top-performing models, however, aggregation accuracy declined sharply as context size increased. Across all models evaluated, average accuracy decreased by 16.5% at 128K context and 25.6% at 200K context relative to 32K performance. On average, aggregation accuracy degraded 2x faster than for single document tasks. Several models that performed well at smaller context sizes experienced severe degradation during aggregation tasks at larger scales, as can be seen in Figure 16.

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	Single Doc Accuracy 32K	Single Doc Accuracy 128K	Accuracy Change (32K -> 128K, pp)	Single Doc 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
Qwen3-4B-Instruct-2507	75.1%	36.0%	-39.2	20.7%	-54.4

Figure 16: *Multi-document Aggregation Tasks – Decrease Across Context Sizes*

Llama-4-Maverick-17B-128E-Instruct, for example, declined from 91% accuracy at 32K context to 49.4% at 128K context. Similarly, GLM-4.7 experienced relatively limited degradation at 128K context before declining substantially at 200K context. GLM-4.6 demonstrated the largest overall reduction in aggregation accuracy, declining by 68.1 percentage points between 32K and 200K context sizes.

Overall, aggregation workloads demonstrated substantially greater sensitivity to increasing context size than single-document retrieval tasks. These findings suggest that cross-document reasoning and information synthesis remain major challenges for long-context LLM systems, even among models that perform strongly on simpler retrieval-oriented workloads.

Hallucination Probing Tasks

The final evaluation category in the RIKER benchmark specifically probes for hallucination behavior. These tasks evaluate whether models incorrectly generate information that does not exist within the provided corpus. One hallucination task asks about entities that are entirely absent from the dataset, while the second asks about optional fields intentionally omitted from specific documents. Correct responses require models to acknowledge the absence of information using responses such as “Unknown” or “N/A”.

Compared to retrieval and aggregation workloads, hallucination probing tasks demonstrated substantially greater stability across increasing context sizes. Qwen3.5-397B-A17B (Thinking) showed the strongest overall hallucination resistance, maintaining greater than 99% accuracy across all evaluated context lengths. Other

top models, such as Qwen3.5-122B-A10B (Thinking), GPT-5.4 (Medium Reasoning), and Kimi-K2.5 (Thinking) similarly maintained strong and highly stable hallucination performance as context size increased.

Although top performing models were found to remain stable, some models still experienced notable degraded accuracy as context length increased. The largest total decrease in accuracy was seen in GLM-4.6, which despite achieving 91% accuracy at the 32K context size, decreased its accuracy by 61.2 percentage points when extended to a 200K context size.

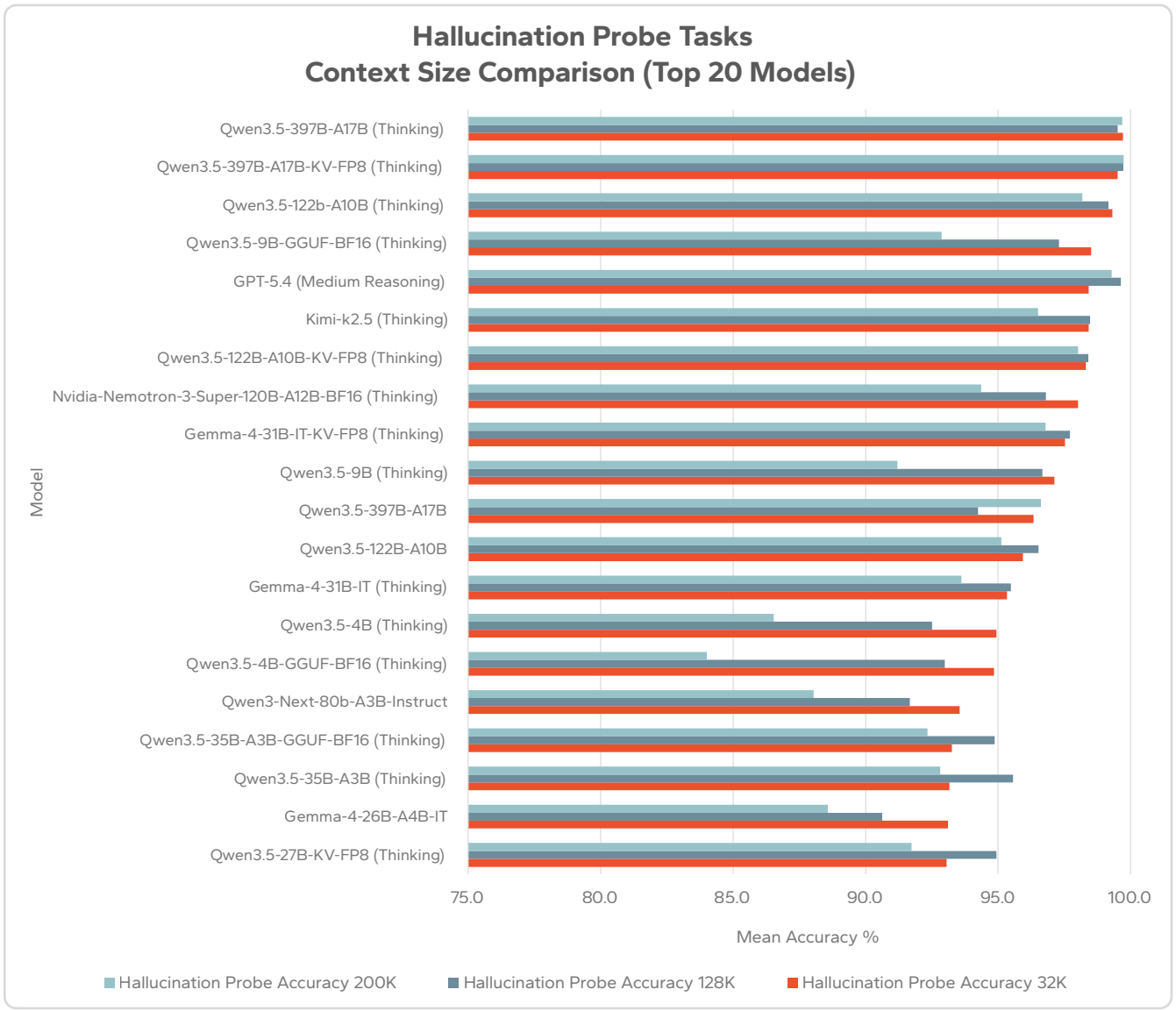


Figure 17: *Hallucination Probing Tasks*

Although certain models experienced significant degradation in hallucination resistance at larger context sizes, average performance declines were comparatively limited relative to both single document and multi-document retrieval tasks. While some models, such as GLM-4.6 demonstrated substantial reduction in hallucination accuracy, many models demonstrated stable or even slightly improved hallucination performance at larger context sizes.

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
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 18: *Hallucination Probing Tasks - Context Size Comparison*

At 128K context, hallucination accuracy decreased by an average of 2.79% across evaluated models, substantially less than the degradation observed in both single-document and aggregation retrieval tasks. Degradation was more severe at the 200K context size, with the average accuracy decreasing by 9.78%. While significant, the loss of accuracy was still lower than in either set of retrieval tasks.

Overall, hallucination resistance appeared less sensitive to context scaling than retrieval-oriented performance. These findings suggest that retrieval degradation and hallucination behavior may represent partially distinct failure modes in long-context LLM systems. While models increasingly struggled to retrieve and aggregate information at larger context sizes, many retained the ability to recognize when information was absent from the provided corpus.

Operational Implications for Enterprise AI

Knowledge retrieval and contextual understanding are core to enterprise AI deployments today. The RIKER findings suggest several operational considerations for organizations deploying retrieval-oriented AI workflows in production environments.

Context window size should not be treated as a proxy for retrieval capability – Although model context windows continue to expand rapidly, most evaluated systems demonstrated measurable retrieval degradation as context size increased. These findings suggest that supported token limits alone are insufficient indicators of retrieval robustness for enterprise workloads.

Multi-document aggregation and synthesis is a key bottleneck – While many models performed strongly on single-document retrieval tasks, aggregation workloads produced substantially greater degradation as context size increased. This is particularly significant because most enterprise retrieval workflows require models to synthesize information distributed across multiple sources rather than retrieve isolated facts from individual documents.

Retrieval stability matters more than peak short-context accuracy – At moderate context sizes, many models achieved similarly strong retrieval performance. As context size increased, however, only a small subset of models maintained stable accuracy. These findings suggest that retrieval robustness across varying workload sizes may be more operationally important than peak performance at smaller context lengths.

Reasoning mode selection can materially impact retrieval quality - Thinking-enabled models consistently outperformed their non-thinking counterparts across multiple model families and context sizes. For enterprise deployments, reasoning configuration may be as important as model selection itself when optimizing retrieval-oriented workflows.

Hallucination resistance may scale differently than retrieval performance – While retrieval and aggregation accuracy consistently declined as context size increased, hallucination probing tasks remained comparatively stable across many models, especially at the 128K context size. This suggests that retrieval degradation and hallucination behavior may represent partially distinct failure modes in long-context LLM systems.

What This Means for the Industry

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

The context window arms race has outpaced enterprise usability. Frontier vendors now advertise 200K, 1M, and 2M token windows as headline features, and procurement teams have begun treating context length as a primary selection criterion. The RIKER data shows that's the wrong question. Of 54 models tested, only 3

sustained better than 95% retrieval accuracy at 200K, and aggregation accuracy collapsed by an average of 25.55% at the same context size across all models. Enterprises evaluating LLMs for RAG, knowledge management, or agentic workflows should be asking vendors for retrieval-stability curves across context scale, not token caps.

The benchmarks the industry relies on are not built for enterprise reality. Public leaderboards increasingly struggle to reflect enterprise reality due to training-data contamination, LLM-as-judge subjectivity, and synthetic tasks that poorly approximate production retrieval workflows. RIKER was designed in direct response: an inverted-generation methodology that produces a coherent, contamination-resistant simulated universe; deterministic grading that removes subjectivity from scoring; and corpora shaped to mirror real enterprise document types, leases, HR policies, field reports, rather than trivia. Evaluating production AI requires evaluation methods built for production AI.

Enterprise AI requires coordinated benchmarks. Enterprise AI is complex and its capability cannot be reduced to a single benchmark dimension. RIKER measures information retrieval and hallucination resistance. KAMI, Signal65 and Kamiwaza's agentic capability benchmark, measures the next layer up, whether models can plan, act, and recover within enterprise workflows. PINNACLE, the next benchmark in the series, will extend the framework to evaluate additional enterprise-oriented performance and quality dimensions. Together, these benchmarks form a deliberately enterprise-grade evaluation suite: contamination-resistant, deterministically graded, and built around the workloads that determine whether AI delivers operational value in production. Signal65 and Kamiwaza will continue to publish results, methodology, and datasets openly, because the buying decisions in front of CIOs today are too consequential to rest on benchmarks designed for demos.

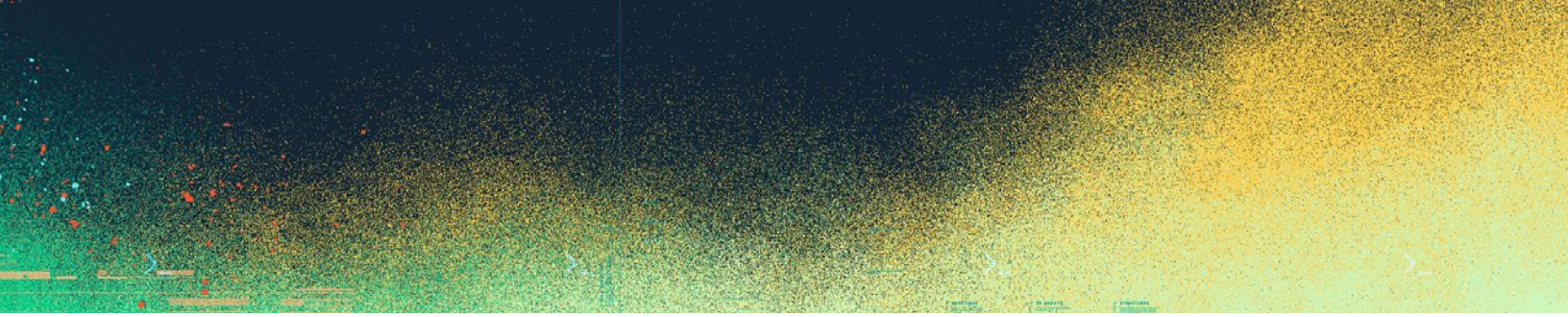
Limitations and Future Work

While the RIKER 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 is additionally focused on a limited set of enterprise-oriented domains, including HR records, commercial leases, and facility field reports. Retrieval performance may differ in highly specialized domains such as healthcare, legal analysis, scientific research, or software engineering workflows that rely on more technical or domain-specific terminology.

Current RIKER 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 continue expanding both the RIKER and KAMI benchmarks to evaluate broader categories of enterprise AI workloads, additional retrieval architectures, and increasingly complex long-context reasoning scenarios. Signal65 and Kamiwaza are further expanding these benchmarking efforts with PINNACLE to further enable organizations to evaluate both model quality and deployment solutions in relation to common enterprise use cases and personas.

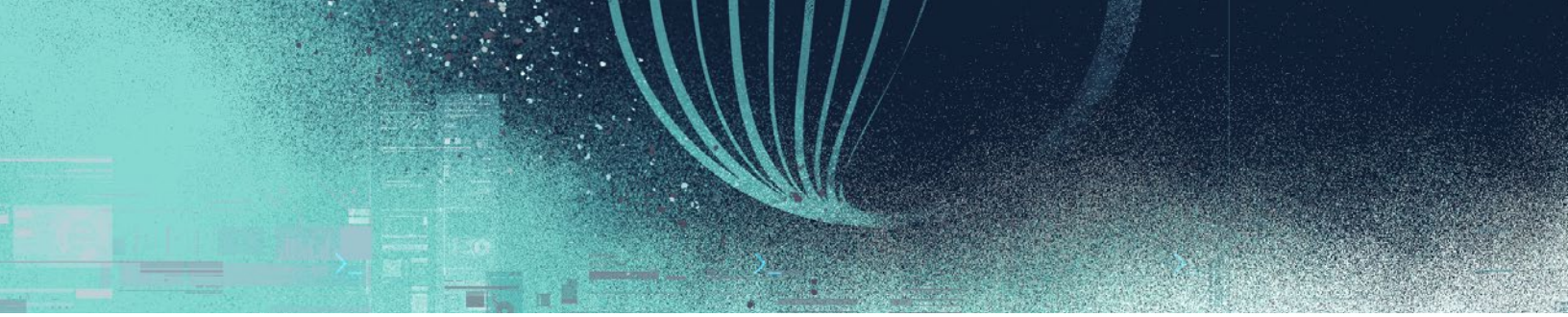


Conclusion

As enterprise AI deployments increasingly rely on retrieval-oriented workflows, accurately evaluating long-context retrieval capability becomes increasingly important. The results of the RIKER benchmark demonstrate that while many modern LLMs perform strongly at moderate context sizes, retrieval accuracy frequently degrades as context length and retrieval complexity increase.

In particular, cross-document aggregation and synthesis workloads remain substantially more challenging than single-document retrieval tasks, and effective long-context retrieval capability varies dramatically across models. These findings suggest that advertised context window size alone is not a reliable indicator of enterprise retrieval performance. The results also indicate that reasoning configuration can materially affect retrieval quality, making deployment settings an important part of enterprise model evaluation.

RIKER provides a framework for evaluating these enterprise-oriented retrieval workloads while reducing susceptibility to contamination and LLM-as-a-judge bias. Future work will continue expanding benchmark coverage and retrieval evaluation capabilities as long-context enterprise AI systems continue to evolve.



Appendix

Overall Results

Model	Overall Mean Accuracy - 32K	Overall Mean Accuracy - 128K	Overall Mean Accuracy - 200K
Qwen3.5-397B-A17B (Thinking)	99.1%	96.0%	95.6%
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	98.9%	96.3%	95.5%
Gemma-4-31B-IT-KV-FP8 (Thinking)	98.4%	95.3%	93.1%
Qwen3.5-122B-A10B (Thinking)	98.1%	95.2%	94.5%
Kimi-K2.5 (Thinking)	98.1%	96.1%	94.4%
GPT-5.4 (Medium Reasoning)	97.9%	96.2%	95.3%
Qwen3.5-122B-A10B-KV-FP8 (Thinking)	97.8%	95.4%	93.9%
Qwen3.5-397B-A17B	97.7%	93.0%	93.3%
DeepSeek-v3.2 (Thinking)	97.4%	95.6%	–
Gemma-4-31B-IT (Thinking)	97.3%	94.3%	91.6%
Qwen3.5-9B-GGUF-BF16 (Thinking)	97.2%	92.1%	88.7%
GLM-4.5	97.1%	86.0%	–
Qwen3.5-122B-A10B	97.1%	94.1%	91.0%
GPT-5 (Medium Reasoning)	97.0%	–	–
Nvidia-Nemotron-3-Super-120B-S12B-BF16 (Thinking)	96.6%	91.4%	86.2%

Qwen3.5-35B-A3B (Thinking)	96.0%	94.4%	92.6%
MiniMax-M2.1	96.0%	84.5%	–
Qwen3.5-9B (Thinking)	95.9%	90.0%	86.3%
Gemma-4-31B-IT	95.9%	91.8%	89.1%
GPT-5 (Low Reasoning)	95.9%	–	–
Qwen3.5-27B-KV-FP8 (Thinking)	95.7%	95.1%	93.3%
Qwen3.5-35B-A3B-GGUF-BF16 (Thinking)	95.7%	94.3%	92.9%
Qwen3.5-4B-GGUF-BF16 (Thinking)	95.4%	90.5%	83.6%
MiniMax-M2	95.3%	88.5%	–
Qwen3.5-27B	95.3%	92.2%	89.6%
Qwen3.5-27B (Thinking)	95.0%	95.3%	93.4%
DeepSeek-v3.2	94.7%	92.5%	–
Qwen3.5-35B-A3B	94.4%	91.3%	87.9%
Gemma-4-26B-A4B-IT-KV-FP8 (Thinking)	94.2%	86.5%	79.0%
Qwen3.5-4B (Thinking)	94.2%	89.8%	83.5%
Qwen3-Next-80B-A3B-Instruct	94.1%	88.0%	81.3%
Qwen3.6-27B	93.5%	91.5%	90.0%
Qwen3.5-9B	93.2%	87.7%	82.5%
Gemma-4-26B-A4B-IT (Thinking)	92.7%	85.1%	76.0%
GLM-4.6	92.6%	84.3%	35.8%
DeepSeek-v3.1	92.2%	89.8%	–
Qwen3-Next-80B-A3B-Instruct-FP8	91.9%	86.2%	78.8%
Gemma-4-26B-A4B-IT	91.6%	82.8%	76.8%

Qwen3.6-35B-A3B (Thinking)	91.2%	87.5%	86.1%
GLM-4.5-Air	91.1%	70.6%	–
Qwen3-Coder-480B-A35B-Instruct-KV-FP8 (Thinking)	90.7%	77.5%	69.8%
Qwen3-Coder-480B-A35B-Instruct-FP8	90.5%	79.0%	72.3%
Qwen3-Coder-480B-A35B-Instruct	90.5%	78.5%	71.7%
GPT-5.4-Mini (Medium Reasoning)	90.0%	84.4%	82.7%
Qwen3.5-4B	89.8%	81.3%	73.6%
Qwen3-235B-A22B-Instruct-2507	89.7%	79.9%	69.5%
Qwen3-Coder-480B-A35B-Instruct (Thinking)	89.7%	77.8%	71.0%
Qwen3.6-27B (Thinking)	89.6%	88.7%	85.9%
GPT-5.4-Mini (High Reasoning)	89.0%	–	–
Qwen3-235B-A22B-Instruct-2507 (Thinking)	88.9%	79.6%	69.6%
Qwen3-235B-A22B-Instruct-2507-KV-FP8 (Thinking)	88.7%	78.9%	68.4%
GPT-5.4 (No Reasoning)	88.7%	–	–
Qwen3-235B-A22B-Instruct-2507-FP8	88.5%	79.8%	69.5%
GPT-5-Chat	87.2%	–	–
GLM-4.7	86.4%	82.9%	57.1%
Llama-4-Maverick-17B-128E-Instruct	85.1%	63.5%	61.1%
Llama-4-Maverick-17B-128E-Instruct-FP8	83.4%	62.7%	57.6%
Llama-3.1-405B-Instruct	83.1%	54.7%	–
Qwen3-32B	82.7%	–	–

Qwen2.5-72B-Instruct	82.5%	–	–
Qwen3-32B-FP8	82.3%	–	–
Qwen3-32B-OTF-FP8	81.4%	–	–
DeepSeek-v3	79.8%	68.1%	–
Qwen2.5-32B-Instruct	79.2	–	–
Qwen3-4B-Instruct-2507	79.1%	59.6%	49.2%
Qwen3-30B-A3B-Instruct-2507	77.5%	68.2%	64.5%
Qwen3-Coder-30B-A3B-Instruct (Thinking)	77.3%	63.1%	55.9%
Qwen3-Coder-30B-A3B-Instruct-KV-FP8 (Thinking)	76.2%	59.7%	55.5%
Qwen3-Coder-30B-A3B-Instruct	75.8%	63.1%	55.4%
Qwen3.5-2B (Thinking)	73.7%	61.8%	52.4%
Qwen2.5-14B-Instruct	72.5%	–	–
Qwen3-14B	71.5%	–	–
Qwen3-14B-FP8	71.0%	–	–
Qwen3-14B-OTF-FP8	70.6%	–	–
Qwen2.5-Coder-14B-Instruct	69.2%	–	–
Llama-3.1-70B-Instruct	68.9%	40.6%	
Llama-4-Scout-17B-16E-Instruct	66.7%	49.4%	45.1%
Llama-3.3-70B-Instruct	66.2%	37.8%	–
Qwen3-8B	64.1%	–	–
Nvidia-Nemotron-3-Super-120B-S12B-BF16	58.8%	46.7%	45.4%
Qwen3.5-2B	58.4%	46.7%	–

Qwen3-4B	54.2%	–	–
Llama-3.1-8B-Instruct	54.1%	42.9%	–
Granite-4.0-H-Small	53.5%	42.8%	–
Qwen2.5-Coder-7B-Instruct	53.3%	–	–
Llama-3.2-3B-Instruct	40.5%	30.1%	–
Qwen3.5-0.8B	39.6%	30.2%	29.4%
Granite-4.0-H-Micro	37.9%	30.4%	–
Granite-4.0-H-Tiny	32.0%	20.2%	–
Llama-3.2-1B-Instruct	24.5%	28.5%	–

Figure 19: *Overall Mean Accuracy*

Single Document Task Results

Model	Single Document Accuracy - 32K	Single Document Accuracy - 128K	Single Document Accuracy - 200K
Qwen3.6-27B (Thinking)	99.7%	97.9%	97.2%
Qwen3.5-397B-A17B	99.1%	93.0%	94.7%
Qwen3.5-397B-A17B (Thinking)	99.1%	95.6%	95.4%
Qwen3.6-27B	99.1%	95.7%	95.6%
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	99.0%	95.6%	95.2%
Qwen3.5-35B-A3B	98.5%	95.1%	93.6%
Qwen3.5-27B	98.4%	93.9%	94.6%
GPT-5.4 (No Reasoning)	98.4%	–	–
Qwen3.6-35B-A3B (Thinking)	98.3%	96.2%	96.0%
Gemma-4-31B-IT-KV-FP8 (Thinking)	98.3%	95.1%	92.2%

MiniMax-M2.1	98.2%	91.2%	–
Gemma-4-31B-IT (Thinking)	98.1%	94.6%	90.8%
Qwen3.5-122B-A10B (Thinking)	97.8%	94.3%	92.8%
Kimi-K2.5 (Thinking)	97.8%	94.8%	94.1%
Qwen3.5-122B-A10B	97.6%	94.1%	93.2%
Qwen3.5-35B-A3B (Thinking)	97.5%	93.3%	92.8%
GLM-4.5	97.4%	90.5%	–
Qwen3.5-122B-A10B-KV-FP8 (Thinking)	97.4%	94.7%	92.7%
Gemma-4-31B-IT	97.3%	93.8%	89.7%
DeepSeek-v3.2	97.2%	92.9%	–
GPT-5.4 (Medium Reasoning)	97.1%	93.5%	92.1%
Qwen3.5-35B-A3B-GGUF-BF16 (Thinking)	97.0%	93.6%	93.1%
Qwen3.5-4B-GGUF-BF16 (Thinking)	97.0%	91.6%	86.9%
MiniMax-M2	96.6%	93.1%	–
Qwen3-235B-A22B-Instruct-2507 (Thinking)	96.6%	86.3%	83.2%
Qwen3.5-9B-GGUF-BF16 (Thinking)	96.6%	91.9%	90.9%
Qwen3-235B-A22B-Instruct-2507	96.4%	86.1%	82.9%
DeepSeek-v3.2 (Thinking)	96.1%	94.9%	–
DeepSeek-v3.1	96.0%	93.4%	–
Qwen3.5-4B	96.0%	86.2%	85.6%
Qwen3.5-9B	96.0%	89.7%	85.6%
Qwen3-235B-A22B-Instruct-2507-FP8	96.0%	86.1%	82.4%

Qwen3.5-27B-KV-FP8 (Thinking)	95.9%	94.4%	93.3%
Qwen3-235B-A22B-Instruct-2507-KV-FP8 (Thinking)	95.7%	85.3%	82.7%
Llama-4-Maverick-17B-128E-Instruct	95.7%	80.7%	76.0%
Gemma-4-26B-A4B-IT (Thinking)	95.6%	84.9%	75.2%
Gemma-4-26B-A4B-IT-KV-FP8 (Thinking)	95.5%	86.5%	75.5%
GPT-5.4-Mini (Medium Reasoning)	95.5%	92.8%	89.4%
Qwen3.5-9B (Thinking)	95.4%	90.0%	89.3%
GLM-4.7	95.3%	87.3%	74.9%
Qwen3.5-27B (Thinking)	95.2%	94.3%	93.0%
Qwen3.5-4B (Thinking)	95.1%	89.7%	85.8%
Qwen3-Next-80B-A3B-Instruct	94.7%	90.9%	91.7%
GPT-5.4-Mini (High Reasoning)	94.7%	–	–
Nvidia-Nemotron-3-Super-120B-S12B-BF16 (Thinking)	94.6%	94.0%	89.8%
GLM-4.6	94.3%	89.8%	53.0%
GPT-5 (Medium Reasoning)	93.4%	–	–
Qwen2.5-72B-Instruct	93.4%	–	–
Qwen3-Next-80B-A3B-Instruct-FP8	93.4%	90.1%	90.6%
Llama-3.1-405B-Instruct	93.3%	67.3%	–
Qwen3-Coder-480B-A35B-Instruct-KV-FP8 (Thinking)	93.0%	81.4%	80.7%
Llama-4-Maverick-17B-128E-Instruct-FP8	92.9%	80.1%	74.6%
Qwen3-Coder-480B-A35B-Instruct (Thinking)	92.6%	81.8%	80.3%

Nvidia-Nemotron-3-Super-120B-S12B-BF16	92.5%	86.5%	85.9%
Gemma-4-26B-A4B-IT	92.4%	82.4%	73.7%
Qwen3-Coder-480B-A35B-Instruct	92.4%	82.5%	81.7%
GPT-5 (Low Reasoning)	91.8%	–	–
Qwen3-Coder-480B-A35B-Instruct-FP8	91.7%	82.6%	82.6%
GLM-4.5-Air	91.2%	76.5%	–
Llama-3.1-70B-Instruct	90.2%	60.6%	–
GPT-5-Chat	89.9%	–	–
Qwen3-Coder-30B-A3B-Instruct	89.3%	81.6%	78.3%
Qwen3-30B-A3B-Instruct-2507	89.1%	80.4%	80.5%
Qwen3-Coder-30B-A3B-Instruct (Thinking)	89.1%	81.4%	79.6%
Qwen3-4B-Instruct-2507	89.0%	73.6%	57.9%
Qwen2.5-32B-Instruct	89.0%	–	–
Qwen3-Coder-30B-A3B-Instruct-KV-FP8 (Thinking)	88.6%	80.1%	78.4%
DeepSeek-v3	88.5%	81.9%	–
Qwen3-32B-FP8	88.5%	–	–
Qwen3-32B	87.8%	–	–
Qwen3-14B-FP8	86.8%	–	–
Qwen3-32B-OTF-FP8	86.3%	–	–
Qwen3-14B-OTF-FP8	85.6%	–	–
Qwen3-14B	85.5%	–	–
Llama-4-Scout-17B-16E-Instruct	84.1%	71.9%	58.4%

Qwen2.5-Coder-14B-Instruct	81.7%	–	–
Llama-3.3-70B-Instruct	81.4%	47.8%	–
Qwen3-8B	80.2%	–	–
Qwen2.5-14B-Instruct	80.0%	–	–
Qwen3-4B	79.8%	–	–
Qwen3.5-2B	76.0%	65.2%	–
Qwen3.5-2B (Thinking)	74.5%	61.3%	54.8%
Qwen3.5-0.8B	71.5%	56.4%	53.8%
Qwen2.5-Coder-7B-Instruct	67.9%	–	–
Llama-3.1-8B-Instruct	67.0%	49.6%	–
Granite-4.0-H-Small	66.6%	48.5%	–
Llama-3.2-3B-Instruct	49.9%	37.8%	–
Granite-4.0-H-Tiny	45.1%	36.4%	–
Granite-4.0-H-Micro	44.4%	29.2%	–
Llama-3.2-1B-Instruct	15.1%	13.5%	–

Figure 20: *Single Document Retrieval Tasks*

Multi-Document Aggregation Tasks

Model	Aggregation Accuracy - 32K	Aggregation Accuracy - 128K	Aggregation Accuracy - 200K
Gemma-4-31B-IT-KV-FP8 (Thinking)	99.4%	93.0%	90.3%
GPT-5 (Medium Reasoning)	99.2%	–	–
Qwen3.6-35B-A3B (Thinking)	98.6%	93.4%	91.6%
Gemma-4-31B-IT (Thinking)	98.4%	92.7%	90.5%

Qwen3.5-397B-A17B (Thinking)	98.4%	93.0%	91.7%
GPT-5.4 (Medium Reasoning)	98.3%	95.5%	94.3%
Qwen3.5-27B (Thinking)	98.2%	96.4%	94.8%
Qwen3.5-27B-KV-FP8 (Thinking)	98.2%	95.9%	94.9%
Qwen3.5-27B	98.1%	91.3%	86.6%
Gemma-4-31B-IT	98.0%	86.0%	83.5%
Kimi-K2.5 (Thinking)	98.0%	95.1%	92.5%
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	98.0%	93.5%	91.6%
Qwen3.6-27B (Thinking)	98.0%	95.6%	93.2%
Qwen3.5-122B-A10B	97.7%	91.6%	84.6%
Qwen3.5-122B-A10B-KV-FP8 (Thinking)	97.6%	93.2%	91.0%
Qwen3.5-397B-A17B	97.6%	91.9%	88.6%
GPT-5 (Low Reasoning)	97.5%	–	–
Qwen3.5-35B-A3B (Thinking)	97.5%	94.2%	92.2%
Nvidia-Nemotron-3-Super-120B-S12B-BF16 (Thinking)	97.3%	83.5%	74.3%
Qwen3.5-122B-A10B (Thinking)	97.3%	92.1%	92.6%
Qwen3.5-35B-A3B-GGUF-BF16 (Thinking)	97.1%	94.3%	92.3%
DeepSeek-v3.2	96.5%	90.7%	–
DeepSeek-v3.2 (Thinking)	96.3%	93.7%	–
Qwen3.5-35B-A3B	96.3%	88.2%	83.3%
Qwen3.5-9B-GGUF-BF16 (Thinking)	96.1%	87.0%	82.3%
Qwen3.6-27B	96.0%	91.8%	89.7%

MiniMax-M2	95.8%	79.8%	–
Gemma-4-26B-A4B-IT-KV-FP8 (Thinking)	95.6%	85.5%	76.0%
DeepSeek-v3.1	95.2%	83.1%	–
Qwen3.5-9B (Thinking)	95.2%	83.4%	78.2%
GLM-4.5	94.9%	73.5%	–
MiniMax-M2.1	94.9%	74.9%	–
Qwen3-235B-A22B-Instruct-2507-FP8	94.3%	76.6%	59.6%
Qwen3-Next-80B-A3B-Instruct-FP8	94.0%	78.4%	61.0%
Qwen3-Next-80B-A3B-Instruct	93.9%	81.3%	64.0%
Qwen3-235B-A22B-Instruct-2507	93.7%	76.4%	58.8%
Qwen3.5-4B-GGUF-BF16 (Thinking)	93.6%	86.9%	79.0%
GPT-5.4-Mini (High Reasoning)	93.6%	–	–
Qwen3-Coder-480B-A35B-Instruct	93.4%	73.5%	60.7%
Qwen3-Coder-480B-A35B-Instruct-FP8	93.4%	72.1%	62.3%
Qwen3.5-9B	93.4%	81.0%	75.8%
GPT-5-Chat	93.3%	–	–
GPT-5.4-Mini (Medium Reasoning)	93.3%	78.3%	75.3%
Qwen3-235B-A22B-Instruct-2507 (Thinking)	93.1%	76.2%	60.5%
Qwen3-Coder-480B-A35B-Instruct-KV-FP8 (Thinking)	92.4%	71.2%	58.9%
Qwen3.5-4B (Thinking)	92.3%	87.1%	78.2%
Qwen3.5-4B	92.1%	76.6%	64.1%

Qwen3-235B-A22B-Instruct-2507-KV-FP8 (Thinking)	92.0%	75.1%	57.9%
Gemma-4-26B-A4B-IT (Thinking)	91.9%	82.9%	73.1%
GLM-4.6	91.8%	78.3%	23.7%
Qwen3-Coder-480B-A35B-Instruct (Thinking)	91.6%	72.8%	61.5%
Llama-4-Maverick-17B-128E-Instruct	91.0%	49.4%	51.2%
Llama-4-Maverick-17B-128E-Instruct-FP8	90.0%	51.2%	47.7%
Gemma-4-26B-A4B-IT	89.3%	75.2%	68.0%
GLM-4.5-Air	86.5%	48.8%	–
Qwen3-Coder-30B-A3B-Instruct (Thinking)	84.8%	52.9%	37.0%
GPT-5.4 (No Reasoning)	83.2%	–	–
Llama-3.1-405B-Instruct	83.2%	31.5%	–
Qwen3-Coder-30B-A3B-Instruct-KV-FP8 (Thinking)	83.2%	49.7%	39.1%
Qwen3-Coder-30B-A3B-Instruct	82.0%	54.1%	37.0%
Qwen3-30B-A3B-Instruct-2507	81.6%	62.6%	55.5%
GLM-4.7	81.2%	74.9%	42.8%
DeepSeek-v3	79.6%	52.8%	–
Qwen3-32B	79.3%	–	–
Qwen3-32B-OTF-FP8	78.8%	–	–
Qwen2.5-72B-Instruct	78.3%	–	–
Qwen3-32B-FP8	77.3%	–	–
Qwen3-4B-Instruct-2507	75.1%	36.0%	20.7%

Qwen2.5-32B-Instruct	70.5%	–	–
Llama-3.1-70B-Instruct	67.0%	19.7%	–
Llama-4-Scout-17B-16E-Instruct	65.6%	28.6%	29.5%
Qwen3.5-2B (Thinking)	65.4%	46.6%	33.8%
Qwen2.5-14B-Instruct	63.1%	–	–
Llama-3.3-70B-Instruct	61.2%	19.5%	–
Nvidia-Nemotron-3-Super-120B-S12B-BF16	53.7%	33.8%	37.1%
Qwen2.5-Coder-14B-Instruct	51.2%	–	–
Qwen3-14B-FP8	47.7%	–	–
Qwen3-14B	45.7%	–	–
Qwen3-14B-OTF-FP8	43.3%	–	–
Qwen3.5-2B	42.6%	21.9%	–
Qwen3-4B	40.3%	–	–
Qwen3-8B	39.1%	–	–
Qwen2.5-Coder-7B-Instruct	37.3%	–	–
Granite-4.0-H-Small	36.2%	14.8%	–
Llama-3.1-8B-Instruct	34.2%	16.8%	–
Granite-4.0-H-Tiny	29.0%	11.3%	–
Qwen3.5-0.8B	21.3%	11.7%	12.1%
Granite-4.0-H-Micro	21.0%	11.0%	–
Llama-3.2-3B-Instruct	16.0%	11.0%	–
Llama-3.2-1B-Instruct	4.8%	2.5%	–

Figure 21: *Multi-document Aggregation Retrieval Tasks*

Hallucination Probing Tasks

Model	Hallucination Probing Accuracy - 32K	Hallucination Probing Accuracy - 128K	Hallucination Probing Accuracy - 200K
DeepSeek-v3.2 (Thinking)	99.7%	98.3%	–
Qwen3.5-397B-A17B (Thinking)	99.7%	99.5%	99.7%
Qwen3.5-397B-A17B-KV-FP8 (Thinking)	99.5%	99.7%	99.7%
Qwen3.5-122B-A10B (Thinking)	99.3%	99.2%	98.2%
GLM-4.5	98.8%	93.8%	–
Qwen3.5-9B-GGUF-BF16 (Thinking)	98.7%	97.3%	92.9%
GPT-5 (Medium Reasoning)	98.7%	–	–
GPT-5 (Low Reasoning)	98.4%	–	–
GPT-5.4 (Medium Reasoning)	98.4%	99.6%	99.3%
Kimi-K2.5 (Thinking)	98.4%	98.5%	96.5%
Qwen3.5-122B-A10B-KV-FP8 (Thinking)	98.3%	98.4%	98.0%
Nvidia-Nemotron-3-Super-120B-S12B-BF16 (Thinking)	98.0%	96.8%	94.4%
Gemma-4-31B-IT-KV-FP8 (Thinking)	97.5%	97.7%	96.8%
Qwen3.5-9B (Thinking)	97.1%	96.7%	91.2%
Qwen3.5-397B-A17B	96.3%	94.2%	96.6%
Qwen3.5-122B-A10B	95.9%	96.5%	95.1%
Qwen3.5-4B-GGUF-BF16 (Thinking)	95.4%	93.0%	84.7%
Gemma-4-31B-IT (Thinking)	95.3%	95.5%	93.6%
GLM-4.5-Air	95.3%	86.3%	–

Qwen3.5-4B (Thinking)	94.9%	92.5%	86.5%
MiniMax-M2.1	94.7%	87.3%	–
Qwen3-Next-80B-A3B-Instruct	93.6%	91.7%	88.0%
MiniMax-M2	93.4%	92.5%	–
Qwen3.5-35B-A3B (Thinking)	93.2%	95.6%	92.8%
Gemma-4-26B-A4B-IT	93.1%	90.6%	88.6%
Qwen3.5-35B-A3B-GGUF-BF16 (Thinking)	93.1%	94.9%	93.3%
Qwen3.5-27B-KV-FP8 (Thinking)	93.1%	94.9%	91.7%
Gemma-4-31B-IT	92.5%	95.6%	93.9%
Qwen3.5-27B (Thinking)	91.8%	95.2%	92.6%
GLM-4.6	91.7%	84.8%	30.5%
Gemma-4-26B-A4B-IT-KV-FP8 (Thinking)	91.5%	87.4%	85.4%
DeepSeek-v3.2	90.5%	93.9%	–
Gemma-4-26B-A4B-IT (Thinking)	90.4%	87.4%	79.6%
Qwen3.5-9B	90.1%	92.3%	86.1%
Qwen3.5-27B	89.4%	91.3%	87.6%
Qwen3-Next-80B-A3B-Instruct-FP8	88.6%	90.1%	84.7%
Qwen3.5-35B-A3B	88.5%	90.6%	86.7%
Qwen3-Coder-480B-A35B-Instruct-KV-FP8 (Thinking)	86.8%	79.8%	69.5%
Qwen3-Coder-480B-A35B-Instruct-FP8	86.6%	82.2%	71.8%
Qwen3-Coder-480B-A35B-Instruct	85.8%	79.5%	72.4%
DeepSeek-v3.1	85.6%	92.6%	–

Qwen3.6-27B	85.6%	87.2%	84.8%
Qwen3-Coder-480B-A35B-Instruct (Thinking)	84.9%	78.9%	71.0%
GPT-5.4 (No Reasoning)	84.1%	–	–
GLM-4.7	82.4%	86.4%	53.3%
Qwen3-14B	81.7%	–	–
GPT-5.4-Mini (Medium Reasoning)	81.5%	82.0%	83.3%
Qwen3-14B-OTF-FP8	81.3%	–	–
Qwen3.5-4B	81.3%	81.1%	70.9%
Qwen3-32B-FP8	80.8%	–	–
Qwen3-32B	80.8%	–	–
Qwen3.5-2B (Thinking)	80.6%	77.4%	68.3%
Qwen3-235B-A22B-Instruct-2507	79.2%	77.1%	66.6%
GPT-5.4-Mini (High Reasoning)	79.1%	–	–
Qwen3-32B-OTF-FP8	79.1%	–	–
Qwen3-235B-A22B-Instruct-2507-KV-FP8 (Thinking)	78.7%	76.2%	64.4%
GPT-5-Chat	78.6%	–	–
Qwen2.5-32B-Instruct	77.6%	–	–
Qwen3-235B-A22B-Instruct-2507 (Thinking)	77.2%	76.5%	65.1%
Qwen3.6-35B-A3B (Thinking)	77.2%	73.0%	70.8%
Qwen3-14B-FP8	77.1%	–	–
Qwen2.5-72B-Instruct	75.7%	–	–
Qwen3-235B-A22B-Instruct-2507-FP8	75.4%	76.6%	66.4%

Qwen2.5-14B-Instruct	74.0%	–	–
Qwen2.5-Coder-14B-Instruct	73.6%	–	–
Qwen3-4B-Instruct-2507	73.1%	68.9%	68.6%
Llama-3.1-405B-Instruct	72.9%	64.9%	–
Qwen3-8B	71.7%	–	–
Qwen3.6-27B (Thinking)	71.6%	72.6%	67.3%
DeepSeek-v3	71.3%	69.5%	–
Llama-4-Maverick-17B-128E-Instruct	69.1%	60.3%	55.7%
Llama-4-Maverick-17B-128E-Instruct-FP8	67.8%	56.5%	50.3%
Qwen3-30B-A3B-Instruct-2507	62.0%	61.5%	57.2%
Llama-3.1-8B-Instruct	60.0%	61.9%	–
Qwen3-Coder-30B-A3B-Instruct (Thinking)	58.3%	55.0%	50.8%
Qwen3-Coder-30B-A3B-Instruct-KV-FP8 (Thinking)	57.1%	49.1%	48.7%
Granite-4.0-H-Small	56.6%	64.7%	–
Qwen3-Coder-30B-A3B-Instruct	56.4%	53.6%	50.6%
Llama-3.3-70B-Instruct	55.9%	45.8%	–
Qwen3.5-2B	55.6%	52.6%	–
Llama-3.2-3B-Instruct	54.2%	41.3%	–
Qwen2.5-Coder-7B-Instruct	53.8%	–	–
Llama-3.2-1B-Instruct	52.6%	69.2%	–
Llama-4-Scout-17B-16E-Instruct	50.3%	47.4%	47.3%
Llama-3.1-70B-Instruct	49.3%	41.3%	–

Granite-4.0-H-Micro	47.3%	50.8%	–
Qwen3-4B	41.9%	–	–
Nvidia-Nemotron-3-Super-120B-S12B-BF16	30.1%	19.6%	13.2%
Qwen3.5-0.8B	24.9%	22.4%	22.1%
Granite-4.0-H-Tiny	21.7%	12.8%	–

Figure 22: *Hallucination Probing Tasks*

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | 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.

IN PARTNERSHIP WITH

KAMI WAZA

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.

signal65

signal65

CONTACT INFORMATION
Signal65 I signal65.com