Tolaga Research
Harness the Power of Intelligence

Analyst Report
November 2026

Longer Context, Broader AI Capabilities

How expanding context windows are transforming AI applications and inference infrastructure

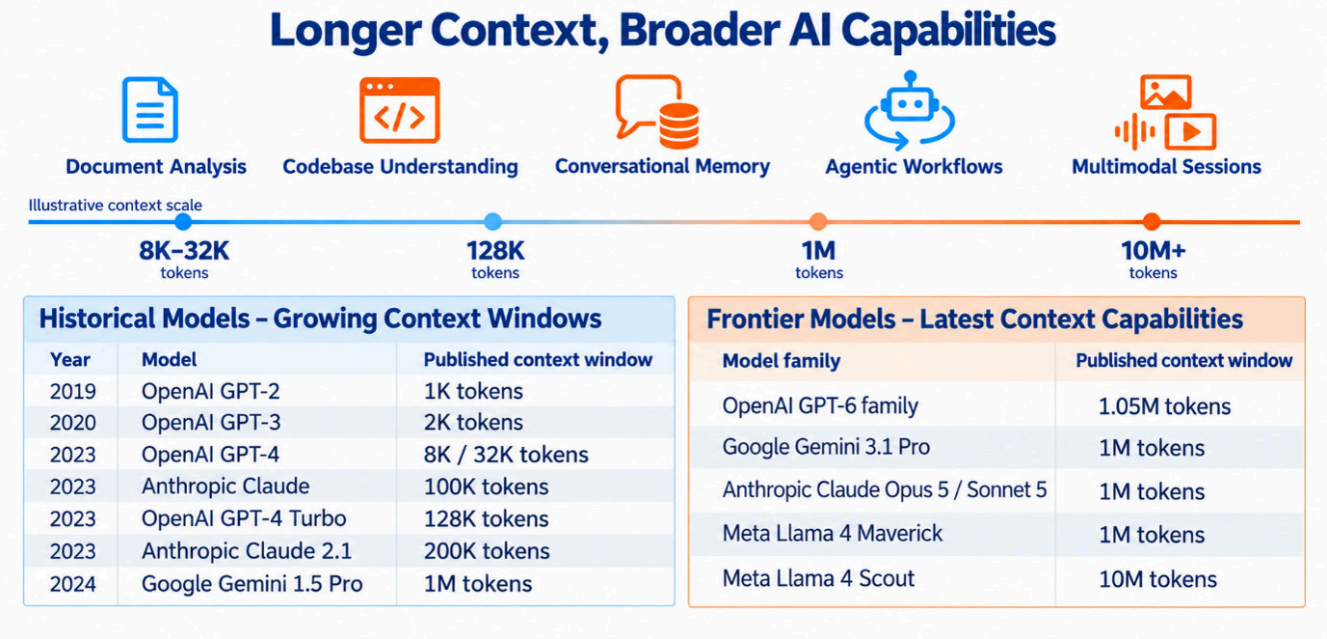
Phil Marshall PhD, Chief Research Officer, Tolaga Research



November, 2026

The Evolution of Context Windows

Context windows have grown by more than three orders of magnitude in just a few years, from a few thousand tokens to over a million. That is enough to accommodate a substantial software repository or several full-length books, greatly expanding the potential applications of large language models (LLMs).



Historical Models – Growing Context Windows

Year	Model	Published context window
2019	OpenAI GPT-2	1K tokens
2020	OpenAI GPT-3	2K tokens
2023	OpenAI GPT-4	8K / 32K tokens
2023	Anthropic Claude	100K tokens
2023	OpenAI GPT-4 Turbo	128K tokens
2023	Anthropic Claude 2.1	200K tokens
2024	Google Gemini 1.5 Pro	1M tokens

Frontier Models – Latest Context Capabilities

Model family	Published context window
OpenAI GPT-6 family	1.05M tokens
Google Gemini 3.1 Pro	1M tokens
Anthropic Claude Opus 5 / Sonnet 5	1M tokens
Meta Llama 4 Maverick	1M tokens
Meta Llama 4 Scout	10M tokens

This rapid expansion also places increased demands on inference infrastructure. Attention, which determines how tokens relate to one another, requires more computation as sequences grow, while the key-value (KV) cache, which stores previously computed representations for reuse, increases memory requirements. Efficient long-context inference therefore depends on balancing computational throughput, memory capacity, and memory bandwidth.

Expanding the Scope of AI Applications

With expanding context windows, AI systems can move beyond one-off exchanges and work with larger, more complex collections of information. A model can analyze a complete report, compare multiple research papers, or examine substantial portions of a software repository within a single request.

In document analysis and software development, this creates opportunities to identify inconsistencies and dependencies spanning multiple sources, such as conflicting terms across contracts or software defects involving several files. In conversational and agentic applications, models can maintain conversation histories, intermediate results, tool outputs, and evolving plans across extended workflows, reducing the need for repeated summarization.

© 2026 Tolaga Research. All rights reserved.

Larger contexts also support more extensive audio, video, and multimodal processing, although token requirements vary considerably by modality. For example, a context window large enough to accommodate an entire novel may hold only a relatively short video, depending on how the content is encoded.

Context Capacity Versus Effective Utilization

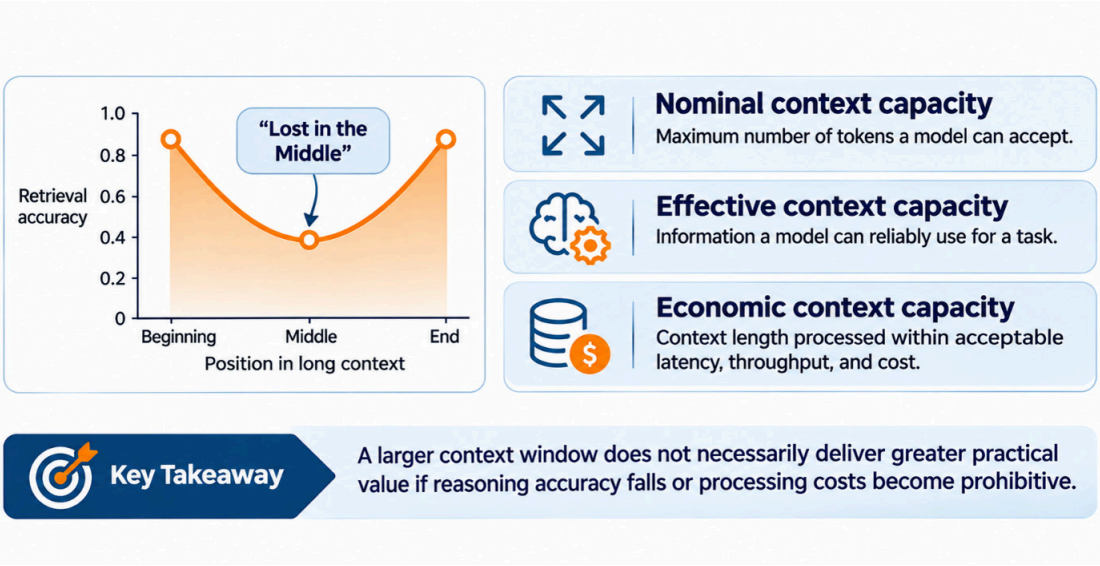
A critical distinction exists between how much information a model can accept and how effectively it can use that information. Greater context capacity does not necessarily translate into better results, particularly when tasks require identifying, combining, or reasoning across information distributed throughout long inputs.

Research has demonstrated these limitations. The *lost in the middle phenomenon*, documented by Liu et al. (2024), describes weaker performance when relevant information appears near the middle of a long sequence rather than near its beginning or end. Retrieving a single fact from a million-token context is also very different from synthesizing evidence, resolving contradictions, or reasoning across multiple dispersed sources.

Three measures of context capacity are therefore useful:

- **Nominal context capacity:** The maximum number of tokens a model can accept.
- **Effective context capacity:** The amount of context a model can reliably retrieve, interpret, and reason over for a particular task.
- **Economic context capacity:** The context length that can be processed within acceptable latency, throughput, and cost constraints.

These measures are interconnected but do not necessarily increase together. A model may support millions of tokens while its reasoning accuracy declines on certain long-context tasks. Similarly, an application may benefit from additional context but become economically impractical if processing time, memory requirements, or infrastructure costs grow excessively.

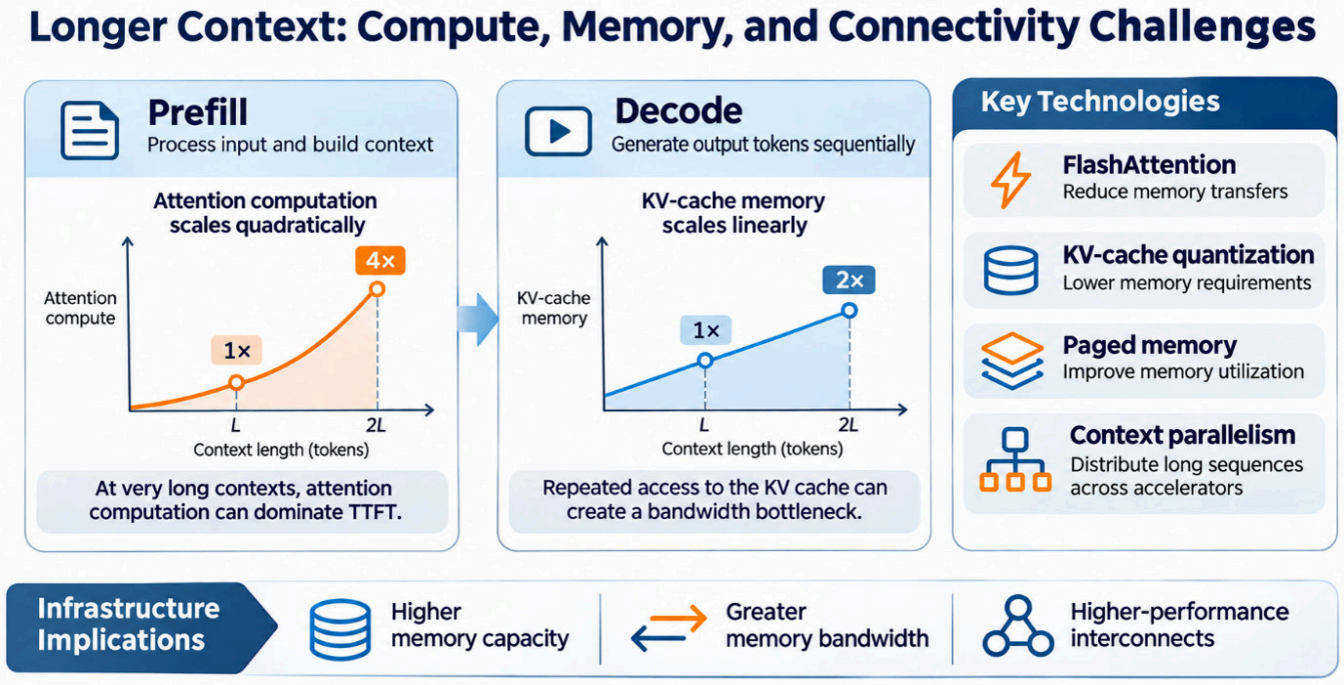


© 2026 Tolaga Research. All rights reserved.

The challenge is therefore not simply to maximize context capacity, but to translate additional context into meaningful AI capabilities without compromising performance or economic efficiency.

The Infrastructure Challenges of Longer Context

Expanding context windows introduces significant computational, memory, and connectivity challenges, particularly during the two principal stages of transformer inference: prefill and decode.



Prefill processes the input sequence and computes the representations needed to begin generating a response. In conventional transformers employing dense attention, attention computation grows approximately with the square of sequence length. Doubling context length can therefore quadruple the attention component of prefill computation, although other operations scale more directly with token count. At very long sequence lengths, attention computation can become a major determinant of time to first token (TTFT).

Decode generates output tokens sequentially, using previously computed key and value representations stored in a KV cache to avoid recalculating them. KV-cache memory requirements grow approximately linearly with context length and the number of concurrent sequences. The stored information must also be accessed repeatedly as output tokens are generated. For long contexts, these repeated memory transfers can create a significant bandwidth bottleneck, limiting token-generation speed even when accelerator computational capacity remains available.

Several technologies help address these challenges. FlashAttention reduces memory transfers during attention processing, while KV-cache quantization lowers memory requirements and associated data movement by storing information at reduced precision. Paged KV-cache management improves memory utilization, while context parallelism distributes long sequences across multiple accelerators, expanding aggregate memory capacity and distributing attention computation.

© 2026 Tolaga Research. All rights reserved.

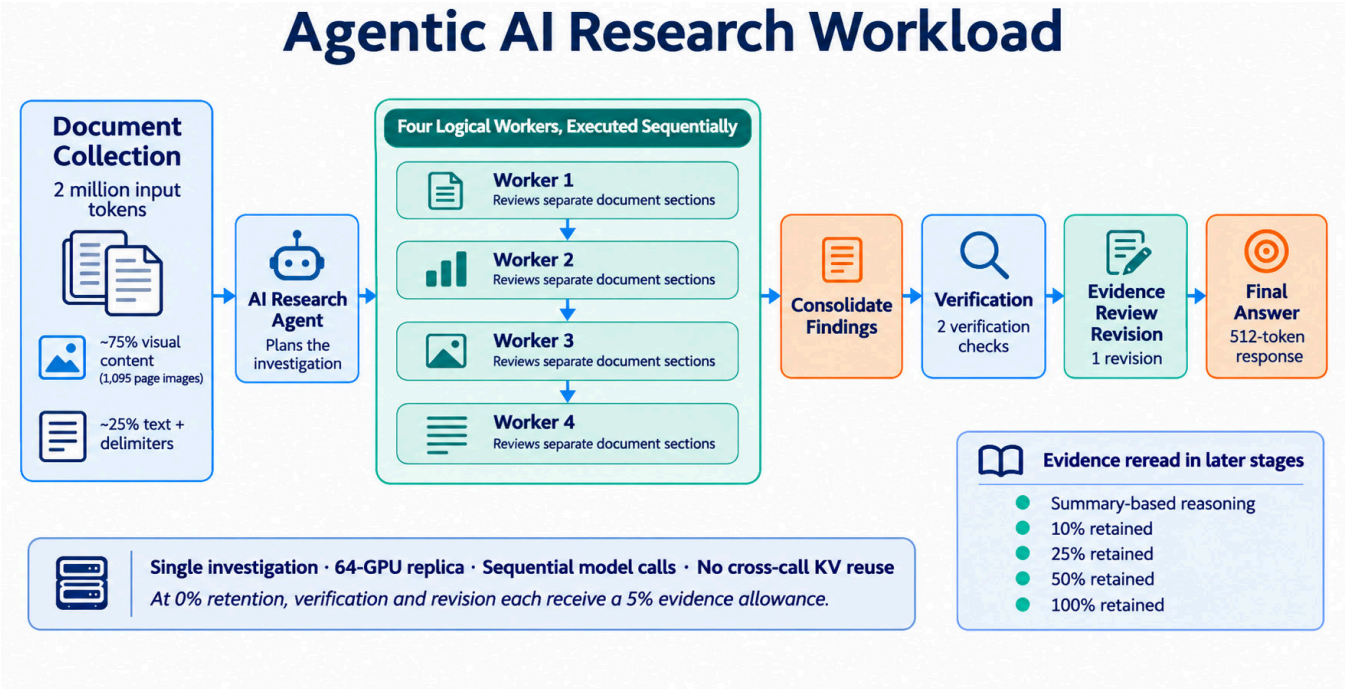
However, context parallelism also introduces communication and synchronization overhead. Its effectiveness depends on how sequences are partitioned, the underlying interconnect architecture, and the balance between computation and data movement. Consequently, distributing longer contexts can improve capacity and performance while increasing the importance of high-bandwidth, low-latency connectivity.

Together, these considerations highlight the importance of matching infrastructure and context-management strategies to the characteristics of the inference workload. The performance implications become particularly significant in agentic applications, where the same source material may be processed repeatedly across multiple stages.

Case Study: Cost of Context for Agentic AI

Tolaga's AI Infrastructure Simulator was used to investigate how large context windows affect agentic AI performance and capacity.

The simulated scenario involves an AI research agent investigating a large document collection containing text, charts, and images. The agent plans the investigation, assigns four workers to analyze separate sections of the collection, consolidates their findings, performs two verification checks, completes one evidence-review revision, and generates a 512-token final response.



The simulation uses a synthetic ~1-trillion-parameter model running on 64 Rubin GPUs. Each investigation processes a source collection of 2 million input tokens, comprising approximately 75% visual content from 1,095 page images and 25% text and delimiters. The analysis varies the proportion of original evidence retained for subsequent processing stages, from summary-based reasoning (0%) to retaining 10%, 25%, 50%, or 100% of the original collection.

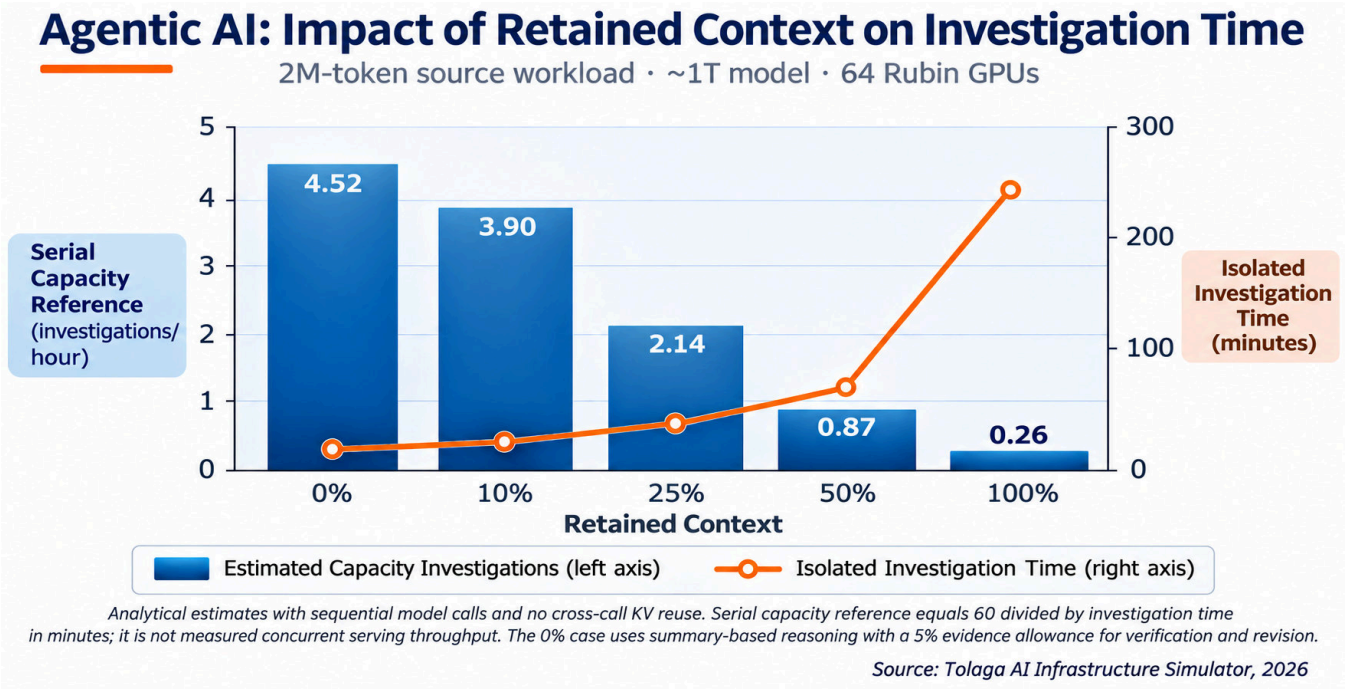
© 2026 Tolaga Research. All rights reserved.

The results demonstrate that increasing the proportion of original evidence retained across successive processing stages from 0% to 100% increases estimated investigation completion time from approximately 13 minutes to 239 minutes, an increase of more than 18-fold. This highlights the substantial performance cost of repeatedly processing large volumes of source material in agentic workflows, rather than relying on consolidated findings from earlier stages.

At 0% retention, subsequent stages primarily rely on worker summaries, with verification and revision calls each receiving an evidence allowance equivalent to approximately 5% of the original source collection. At 100% retention, later-stage calls process approximately two million source tokens, together with instructions and intermediate reports. These represent analytical workload assumptions rather than validated context-processing capabilities of a deployed model.

The analysis considers a single investigation on one 64-GPU replica, with sequential model calls, no competing workloads, and no cross-call KV cache reuse. Although the four workers divide the investigation into separate tasks, they execute sequentially rather than concurrently. The results therefore establish a serial execution baseline that isolates the performance implications of repeated evidence processing. Concurrent scheduling and compatible prefix caching could change both overall completion times and the relative performance penalty associated with retaining larger volumes of evidence.

The simulation estimates execution time and resource requirements. It does not evaluate answer quality or whether retaining additional evidence improves the investigation's conclusions.



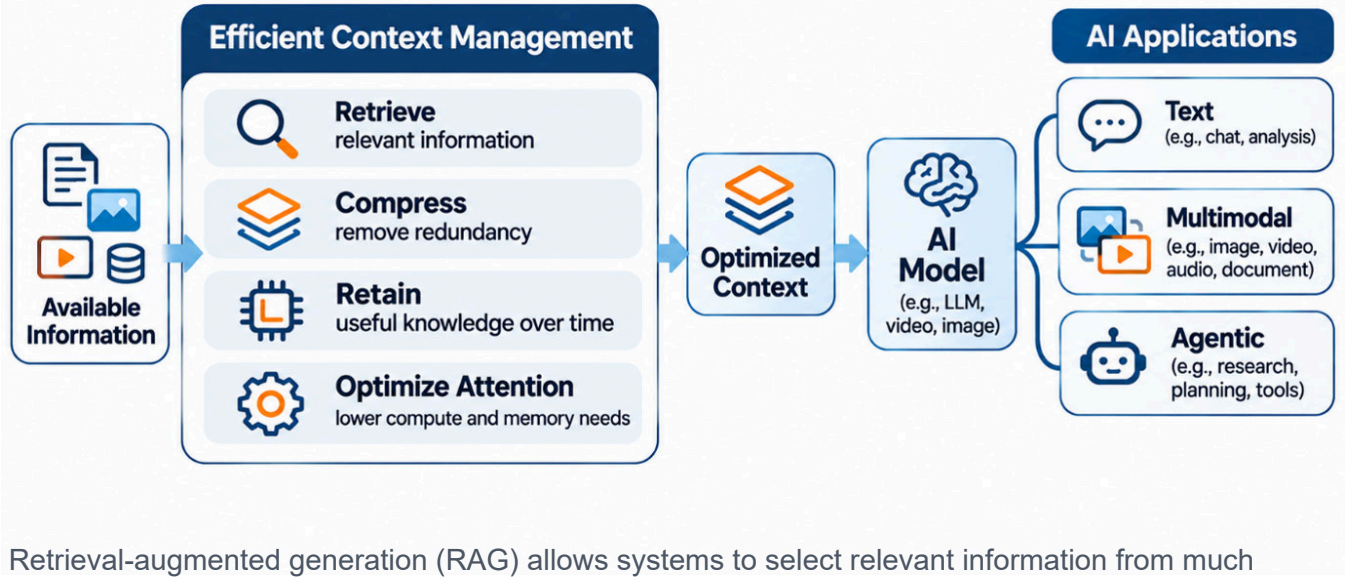
Analytical estimates with sequential model calls and no cross-call KV reuse. Serial capacity reference equals 60 divided by investigation time in minutes; it is not measured concurrent serving throughput. The 0% case uses summary-based reasoning with a 5% evidence allowance for verification and revision.

Source: Tolaga AI Infrastructure Simulator, 2026

© 2026 Tolaga Research. All rights reserved.

Toward More Efficient Context Management

Increasing maximum context length is unlikely to remain the most efficient solution for every workload. Instead, AI systems will increasingly combine expanded context windows with mechanisms that determine which information should be retained, retrieved, compressed, or discarded.



Retrieval-augmented generation (RAG) allows systems to select relevant information from much larger external knowledge bases. Context compression reduces redundant information, while persistent memory mechanisms maintain useful knowledge across interactions without repeatedly processing complete histories. Prefix caching can also reduce redundant prefill computation when multiple requests reuse the same input context. More efficient attention mechanisms further reduce the computational and memory demands of long sequences.

The economic benefits can be substantial, particularly for agentic workflows involving repeated inference operations. Consider an investigation that uses a 100,000-token context across 20 sequential model requests. Reprocessing the complete context at every step would involve approximately two million input tokens across those requests, before accounting for generated outputs. Selective retrieval, compression, and caching can reduce redundant processing, although the savings depend on how much information can be reused and the associated infrastructure implementation.

These approaches suggest a shift from simply maximizing context size toward optimizing the information available to a model at each stage of processing. This is particularly important for agentic and multimodal applications, where the total available information may greatly exceed what is relevant to an individual inference operation.

© 2026 Tolaga Research. All rights reserved.

Strategic Implications

Long-context AI is becoming a systems problem spanning model design, memory, networking, and application architecture. The challenge is no longer simply to accommodate more tokens, but to process increasingly diverse information workloads accurately and economically.

Model developers must improve reasoning reliability across long sequences while limiting the computational and memory demands of attention. Infrastructure vendors face growing pressure to deliver greater usable memory capacity, higher memory bandwidth, and more efficient data movement. As context parallelism becomes more important, high-performance interconnects can also play a larger role in determining system efficiency.

Application developers must decide which information belongs in the context window, what can be retrieved on demand, and what can be retained or reused between inference operations. These decisions directly influence latency, throughput, and infrastructure costs.

For service providers, inference workloads are becoming increasingly diverse, ranging from short conversational exchanges to million-token research tasks. These workloads can place very different demands on computation, memory, and networking, even when using the same model. Consequently, infrastructure efficiency increasingly depends on matching compute, memory, and network architectures to workload requirements rather than maximizing raw accelerator performance.

The economic value of longer context windows will increasingly depend on the amount of useful information processed per unit of infrastructure cost, not simply the maximum number of tokens supported.

Conclusions

Longer context windows are expanding the scope of AI applications, enabling more sophisticated information synthesis, extended interactions, and multimodal processing. However, their practical value depends on reasoning reliability, effective context management, and infrastructure capable of handling increasingly demanding workloads.

As applications become more complex, competitive advantage will increasingly depend on how effectively AI systems combine model capabilities, information management, and infrastructure efficiency.

Bottom line:

The next frontier in long-context AI is not simply processing more information, but **extracting greater value** from the information processed.



© 2026 Tolaga Research. All rights reserved.

About the Author: Phil Marshall, PhD

Phil Marshall's work in AI began more than 30 years ago with Kohonen maps, discrete neural networks, machine learning, and fuzzy logic. More recently, his work has targeted AIenabled solutions for the automotive, healthcare, manufacturing, and smart city sectors. Over the past decade, he has applied natural language processing and large language models to analyze online content and generate distinctive, datadriven industry research insights.

Marshall also brings deep experience in telecom, wireless, mobile, and networking technologies, □elds he has worked in since the early 1990s. More recently, he has combined this infrastructure background with advanced software and modeling capabilities to develop Tolaga's AI Infrastructure Simulator, which is used to analyze the performance, cost, and scalability of emerging AI architectures.

Marshall is a dual citizen of the United States and New Zealand and currently resides in Auckland, New Zealand.

Contact: philip.marshall@tolaga.com

© 2026 Tolaga Research. All rights reserved.