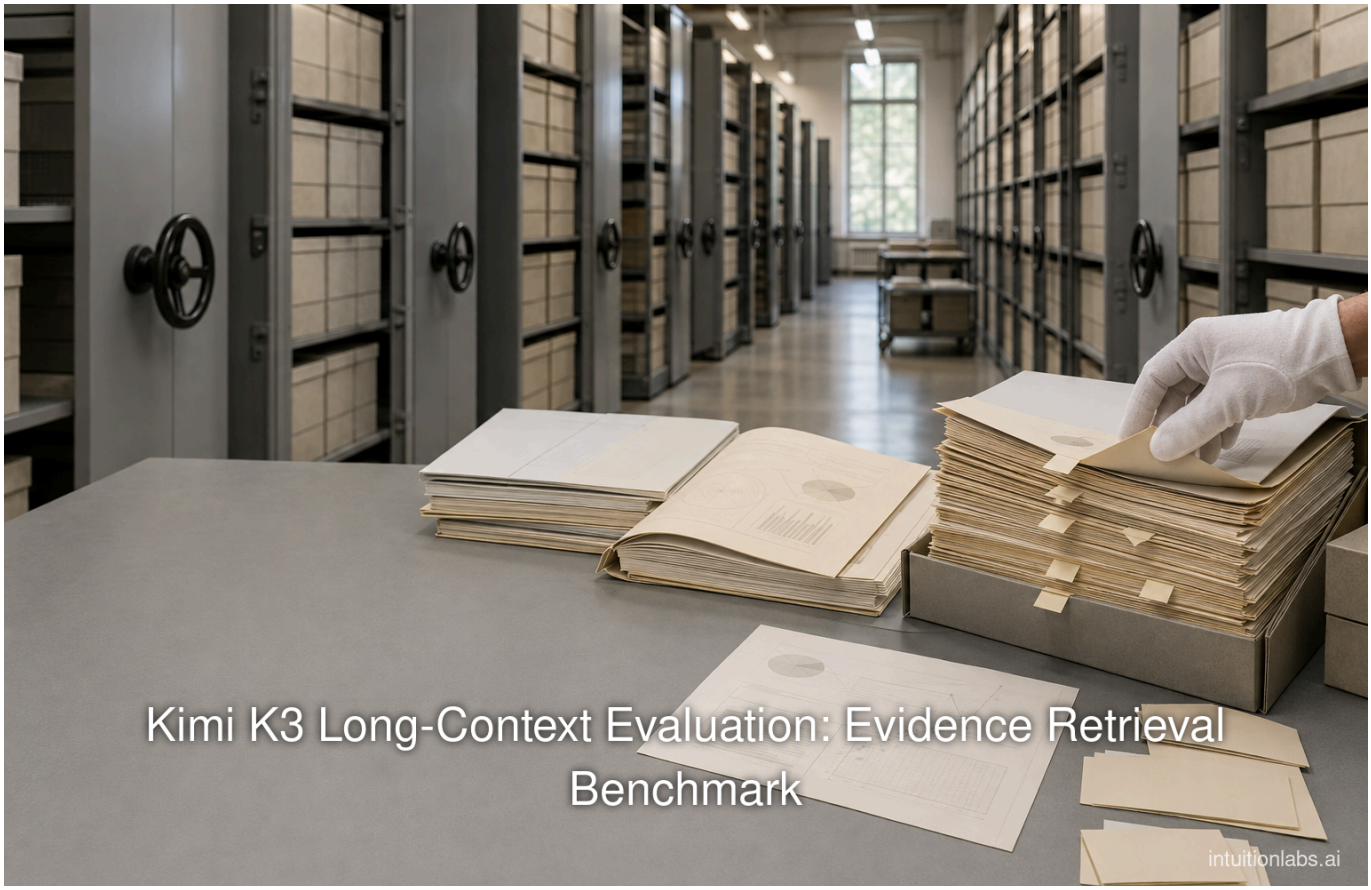
Kimi K3 Long-Context Evaluation: Evidence Retrieval Benchmark

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · kimi k3 · moonshot ai · long context llm · needle in a haystack · context window comparison · AA-LCR benchmark · evidence retrieval AI · LLM benchmark 2026



Executive Summary

Kimi K3, the flagship open-weight large language model (LLM) Moonshot AI introduced and made available before its stated July 27, 2026 full-weights-release deadline (www.kimi.com), ships with a context window of **1,048,576 tokens** (roughly 1 million), a figure independently confirmed by the third-party evaluator Artificial Analysis (artificialanalysis.ai) as well as stated in Moonshot's own **2.8 trillion parameter**, 104 billion activated parameter technical report (arxiv.org). The model's long context is engineered through **Kimi Delta Attention (KDA)** and **Attention Residuals**, a hybrid architecture that interleaves efficient linear-attention layers with periodic full-attention layers to preserve information flow across very long sequences (arxiv.org)(arxiv.org). This report's central, reproducible finding is a gap: Moonshot has not published, and no independent evaluator was found to have published, a standardized needle-in-a-haystack, RULER, or LongBench score for Kimi K3, meaning the exact-fact retrieval-at-depth claim implied by its headline context window remains unverified against the industry's original long-context stress test as of September 2026.

What Moonshot does publish are two retrieval-adjacent figures. On **BrowseComp**, an agentic evidence-gathering benchmark, K3 scores 90.4 percent using its full, uncompacted 1 million token window and a separately reported 91.2 percent using a context-compaction strategy triggered at 300,000 tokens (arxiv.org)(www.kimi.ai), a rare, specific data point suggesting managed context can match or exceed raw window size on at least one retrieval task. On **AA-LCR** (Artificial Analysis Long Context Reasoning), an independently run benchmark whose scores Moonshot explicitly attributes to Artificial Analysis rather than presenting as self-measured (github.com), K3's reported score of 74.7 edges out every comparison model Moonshot selected, including GPT-5.5 (74.3) and Claude Fable 5 (70.0). Independent testing by Artificial Analysis separately places K3 at 57 on its Intelligence Index (artificialanalysis.ai), first on AutomationBench-AA (artificialanalysis.ai), .

On context window size alone, Kimi K3 is no longer a differentiator: OpenAI's GPT-6 Astra (1,050,000 tokens) (developers.openai.com), Anthropic's Claude Opus 5, Sonnet 5, and Fable 5.1 (1M tokens each) (platform.claude.com), Google's Gemini 3.1 Pro and 3.8 Flash (1,048,576 tokens) (ai.google.dev)(ai.google.dev), DeepSeek's V4 family (1M tokens) (api-docs.deepseek.com), and Alibaba's Qwen3.8-Max (1M tokens, plus a 10 million token Qwen-Long variant) (www.alibabacloud.com)(www.alibabacloud.com) have all converged on approximately the same window in 2026. Kimi K3's API is priced at $3.00 per million input tokens, $0.30 on a cache hit, and $15.00 per million output tokens, with flat pricing regardless of context length (artificialanalysis.ai), and it is licensed under a custom Kimi K3 License: a separate agreement is required only when the licensee or an affiliate operates a Model as a Service business and their aggregate revenue exceeds $20 million over any consecutive 12 months (github.com). GitHub made K3 generally available inside GitHub Copilot on August 6, 2026 (github.blog), a documented third-party adoption signal, while Moonshot separately pursues distribution deals with major cloud providers and has filed for a Hong Kong IPO (www.siliconrepublic.com)(www.scmp.com). For research and life-sciences document-analysis buyers, the practical conclusion is that context-window size should no longer drive vendor selection; independently verified retrieval methodology, of the kind this report catalogs and finds still incomplete for Kimi K3, should guide vendor selection.

Introduction and Background

This report examines what can be verified, from primary sources, about the long-context and evidence-retrieval performance of **Kimi K3**, the flagship large language model (LLM) that Moonshot AI, a Beijing-based AI developer, introduced and made available before its stated July 27, 2026 full-weights-release deadline (www.kimi.com). Kimi K3 is a **2.8 trillion parameter** Mixture-of-Experts (MoE) model, of which **104 billion parameters** are activated per token (arxiv.org), and it ships with a context window of **1,048,576 tokens** (commonly rounded to "1 million tokens"). A context window is the maximum amount of text, measured in tokens (sub word units of roughly three to four English characters), that a model can consider at once when generating a response. A "needle in a haystack" (NIAH) test is the industry's original long context stress test, in which a single fact is buried inside a long distractor document and the model must retrieve it. This report treats "evidence retrieval" in the sense relevant to research, legal, and life sciences document analysis: the ability to locate, cite, and reconcile specific facts scattered across one or more long source documents, rather than simply summarizing a text.

IntuitionLabs previously published a technical analysis of Kimi K3's predecessor, Kimi K2, which shipped with a **128,000 token** context window at its July 2025 launch, according to that report (intuitionlabs.ai). That article covers K2's architecture and coding benchmarks; it predates K3, K3's 1 million token window, and the methodology gap this report identifies. This report is a standalone, dated companion: it covers Kimi K3 specifically, traces every long

context benchmark figure Moonshot has published to its origin, compares K3's context window and pricing against the current (September 2026) flagship models from OpenAI, Anthropic, Google, DeepSeek, and Alibaba, and states plainly which standard long context benchmarks (RULER, LongBench, and, for K3 specifically, needle in a haystack testing) have not been published for this model as of the observation date. Every figure below carries the date it was observed or the "as of" date stated by its source, because context windows, prices, and leaderboard standings in this market shift within weeks.

Kimi K3 Architecture and the Engineering Behind the 1 Million Token Window

Kimi K3's long context is not simply "more tokens" bolted onto a standard transformer. Moonshot's technical report, published to arXiv on July 27, 2026, describes a hybrid architecture built around two named mechanisms (arxiv.org). The first, **Kimi Delta Attention (KDA)**, is a linear attention design that "provides efficient long-sequence mixing" and is interleaved with periodic full-attention layers (arxiv.org). The model totals 93 layers, split between 69 KDA layers and 24 Gated Multi-head Latent Attention (MLA) layers plus one dense layer, per Moonshot's published specification. The second mechanism, **Attention Residuals (AttnRes)**, "allows each layer to selectively attend to representations from all preceding layers" (arxiv.org), which Moonshot frames as a way to preserve information flow across very long sequences rather than letting it degrade layer by layer.

Unlike Kimi K2, K3 applies **No Position Encoding (NoPE)** across its attention layers, encoding positional information implicitly through KDA's recurrent gating behavior rather than through an explicit positional embedding (arxiv.org). Moonshot reports this design was reached through a staged context length curriculum during training: the effective window "grows from 8K to 64K tokens during pre-training, and from 256K to 1M tokens" in a later cooldown phase (arxiv.org), using synthetic training tasks constructed so they "can be solved only by attending to information scattered across the full 1M-token context" (arxiv.org) rather than by memorizing local patterns. K3 also ships native vision through MoonViT-V2, "a 27-layer vision transformer with roughly 0.4B parameters" trained from scratch on next-token prediction rather than the contrastive objectives common to earlier vision encoders (arxiv.org).

Moonshot describes K3, in both its GitHub repository and its arXiv paper, as "the world's first open 3T-class model" (github.com), a vendor characterization based on total parameter count rather than an independently audited ranking. The weights and code are released under a permissive but conditional **Kimi K3 License**, under which "both the code repository and the model weights are released under the Kimi K3 License" (raw.githubusercontent.com). If the licensee or any affiliate operates a defined "Model as a Service" business and the aggregate revenue of the licensee and its affiliates exceeds $20 million over any consecutive 12 months, the licensee must enter into a separate agreement with Moonshot before commercial use of the software or derivatives (github.com); a second, independent threshold requires any product built on the software that reaches more than 100 million monthly active users or more than $20 million in monthly revenue to display the "Kimi K3" name "prominently" in its interface (huggingface.co), unless the deployment is purely internal or routed through Moonshot's own certified inference partners.

Methodology: How Long Context and Evidence Retrieval Performance Is Actually Measured

To evaluate a claim like "1 million token context window," a reader needs to know what was tested, not only the maximum window size a vendor advertises. Two evaluation families are most relevant to evidence retrieval and research document analysis: needle-style retrieval tests, which check whether a model can find a fact planted at an arbitrary depth inside a long document, and multi-document reasoning tests, which check whether a model can combine or reconcile facts spread across several documents. OpenAI's own Multi-Round Co-reference Resolution (MRCR) benchmark is one published example of the second family: it inserts several near-identical "needles" into a conversation and asks the model to resolve which one a later reference points to, at context lengths up to 256,000 tokens (openai.com). OpenAI reports that its GPT-5.2 Thinking model achieves "near 100% accuracy on the 4-needle MRCR variant" but only 77.0 percent on the harder eight needle variant within the "128k" to "256k" token range (openai.com), versus 29.6 percent for its immediate predecessor, illustrating that needle count and context length are two separate difficulty dials, and that a single headline context window figure says little about retrieval accuracy under either dial.

Moonshot's own technical report and blog post for Kimi K3 do not report scores on RULER, LongBench, LongBench v2, or MRCR, and do not publish a needle in a haystack accuracy by depth table of the kind several vendors have published historically for their own long context models. Research conducted for this report also found no third party publication of a needle in a haystack result for Kimi K3 at 500,000 or 1,000,000 tokens as of the September 2026 observation date. This is a genuine, dated gap in the public record, not evidence that the model fails such tests: it means the retrieval accuracy at depth claim implied by a "1M token context window" has not yet been independently benchmarked the way NIAH testing was designed to verify, and it is the specific reproducibility gap this report is documenting rather than filling with an invented number.

What Moonshot does report, and what independent evaluators have separately measured, are two retrieval-adjacent benchmarks. The first is **BrowseComp**, an agentic web browsing and evidence gathering benchmark, which Moonshot evaluated under two context management regimes: using the model's full 1 million token context window with no compaction, Kimi K3 scores 90.4 percent (arxiv.org), a figure repeated on Moonshot's own tech blog (www.kimi.ai). The second is **AA-LCR** (Artificial Analysis Long Context Reasoning), version 1.1, an independently run long context reasoning benchmark maintained by the third party evaluator Artificial Analysis; Moonshot's own GitHub documentation explicitly attributes the scores rather than presenting them as self-measured: "Scores are cited from Artificial Analysis as of July 23, 2026" (github.com). Artificial Analysis states its own benchmark runs, including AA-LCR, are conducted as an "Independent test run by Artificial Analysis on dedicated hardware" (artificialanalysis.ai) and separately confirms, on its own model page, that Kimi K3 (max) carries "a context window of 1.0M tokens" (artificialanalysis.ai), corroborating Moonshot's figure from a source with no commercial stake in the result. Because Moonshot selects which competing models appear alongside K3 in its own comparison tables, and because AA-LCR is the only independently sourced long context score Moonshot publishes, this report treats the AA-LCR figures below as vendor-selected but independently measured: the comparison set is Moonshot's choice, while the underlying numbers originate from a third party rather than Moonshot's own test harness.

Benchmark Results: Kimi K3 on Long Context and Retrieval Adjacent Tasks

Table 1 below summarizes the AA-LCR figures Moonshot published for Kimi K3 alongside the specific comparison models Moonshot selected, all sourced to Artificial Analysis as of July 23, 2026 (github.com).

Model	AA-LCR v1.1 score	Comparison source
Kimi K3 (max)	**74.7**	Vendor table citing Artificial Analysis
GPT-5.5 (xhigh)	74.3	Vendor table citing Artificial Analysis
GPT-5.6 Sol (max)	73.7	Vendor table citing Artificial Analysis
GLM-5.2 (max)	71.3	Vendor table citing Artificial Analysis
Claude Fable 5 (max, with fallback)	70.0	Vendor table citing Artificial Analysis
Claude Opus 4.8 (max)	67.7	Vendor table citing Artificial Analysis

On this single independently sourced long context reasoning benchmark, Kimi K3's reported score edges out every model Moonshot chose to compare it against, including GPT-5.5 and Claude Fable 5. This is the strongest available evidence, from a non-Moonshot originator, that K3's long context reasoning is competitive with 2026's leading proprietary systems on at least one standardized test; it is not evidence about needle-style retrieval accuracy specifically, since AA-LCR measures reasoning over long context rather than exact-fact retrieval at depth, and it reflects a comparison set Moonshot selected rather than an exhaustive leaderboard.

On BrowseComp, Moonshot reports a second figure of 91.2 alongside the 90.4 percent uncompacted score, using a context compaction strategy triggered once the conversation exceeds 300,000 tokens (arxiv.org)(www.kimi.ai). That the compacted run scores marginally higher than the uncompacted, full context run is a specific and useful finding for evidence retrieval architecture: it suggests that, for at least this task, actively managing and pruning context can match or slightly exceed simply keeping everything in the window, which cuts against the assumption that a larger raw context window is strictly better for retrieval-style tasks.

Independent evaluation from Artificial Analysis, published July 17, 2026, adds four further data points beyond Moonshot's own disclosures. Kimi K3 "scores 57 on the Artificial Analysis Intelligence Index" (artificialanalysis.ai), a composite benchmark Artificial Analysis places as comparable to Claude Opus 4.8 and GPT-5.5 but behind Claude Fable 5 and GPT-5.6 Sol. On GDPval-AA v2, an agentic task Elo benchmark, K3 "reaches an Elo rating of 1668" (artificialanalysis.ai), and it "takes the #1 position on AutomationBench-AA" with a score of 53 percent (artificialanalysis.ai), a benchmark of automated SaaS workflow completion. On AA-Briefcase, a long horizon knowledge work benchmark, K3 "reaches an overall Elo of 1547" (artificialanalysis.ai), second only to Claude Fable 5 among the models Artificial Analysis tested. On efficiency, Artificial Analysis found K3 required "21% fewer output tokens than K2.6" (artificialanalysis.ai) to reach its higher scores, at an average cost per Intelligence Index task the evaluator calls "similar to GPT-5.6 Sol" at $1.04 per task (artificialanalysis.ai), though Artificial Analysis separately flags the model's per-token API price as "particularly expensive when comparing to other open weight models of similar size" (artificialanalysis.ai).

Context Window and Pricing Comparison: Kimi K3 Against Current Frontier Models

Readers searching for "Kimi K3 vs GPT-4 long context" are, as of September 2026, asking about a comparison point OpenAI has since superseded more than once: GPT-4's 2023-era context window is no longer the reference frontier figure OpenAI publishes pricing or benchmark data against. This report instead compares Kimi K3 against OpenAI's current flagship, GPT-6 Astra, and against the current flagship models from Anthropic, Google, DeepSeek, and Alibaba, each fetched and verified directly from vendor documentation on September 5, 2026.

Table 2 compares the maximum context window and list pricing for the current flagship model in each major family.

Model (vendor)	Max context window	List pricing detail	Source
Kimi K3 (Moonshot AI)	**1,048,576 tokens**	$3.00 input / $15.00 output per 1M tokens; $0.30 per 1M on a cache hit	(artificialanalysis.ai)
GPT-6 Astra (OpenAI)	1,050,000 tokens (922,000 max input)	2x input / 1.5x output surcharge above 272K input tokens	(developers.openai.com) (developers.openai.com)
Claude Opus 5 / Sonnet 5 / Fable 5.1 (Anthropic)	1M tokens each	Not disclosed on the page fetched for this report	(platform.claude.com)
Gemini 3.1 Pro Preview (Google)	1,048,576 tokens	Priced separately from the Flash tier below	(ai.google.dev)
Gemini 3.8 Flash (Google)	1,048,576 tokens	$0.75 input / $3.75 output per 1M tokens (through Dec 31, 2026)	(ai.google.dev)(ai.google.dev)
DeepSeek V4 Pro / V4 Flash (DeepSeek)	1M tokens (384K max output)	Rate not extracted for this report; see source	(api-docs.deepseek.com)
Qwen3.8-Max (Alibaba)	1M tokens	Rate not extracted for this report; see source	(www.alibabacloud.com)(qwen.ai)
Qwen-Long (Alibaba)	10M tokens	Purpose built for multi document review	(www.alibabacloud.com)

Table 2 shows that, on raw context window size, Kimi K3's 1,048,576 token window is no longer a differentiator: it is now the industry baseline among frontier models released or updated in 2026, matched almost exactly by OpenAI's GPT-6 Astra at 1,050,000 tokens (developers.openai.com), Anthropic's entire current Claude 5 family down to Fable 5.1 (platform.claude.com), Google's Gemini 3.1 Pro and 3.8 Flash (ai.google.dev)(ai.google.dev), and DeepSeek's V4 family (api-docs.deepseek.com). Alibaba's Qwen3.8-Max, a 2.4 trillion parameter model released August 3, 2026 (qwen.ai), matches the 1M token figure (www.alibabacloud.com), and Alibaba additionally offers a specialized "qwen-long" endpoint with a 10 million token window explicitly aimed at multi document review tasks (www.alibabacloud.com), ten times the window any other vendor in this table publishes. On pricing, Kimi K3's list rate of $3.00 per million input tokens and $15.00 per million output tokens sits well above Google's Gemini 3.8 Flash rate of $0.75 and $3.75 (ai.google.dev), though Google's figure prices a smaller, faster model rather than a direct architectural peer; Anthropic's and DeepSeek's current per-token rates were not published on the specific pages fetched for this report and are noted as unavailable rather than estimated.

Kimi K3 does differentiate on access model: its documentation states the API "uses flat pay-as-you-go pricing" (platform.kimi.ai) with no separate rate tier for longer contexts, and context caching is automatic, requiring "no cache ID, TTL, or extra parameter" from the developer (platform.kimi.ai), unlike competitors that tier pricing by context length. Access requires only "a successful top-up (minimum $1)" (platform.kimi.ai) against an OpenAI-compatible endpoint using the model ID `kimi-k3`.

Analysis of Key Segments: Licensing and Deployment

Kimi K3's open-weight release does not mean unrestricted commercial use, and the terms above (see Architecture) sit alongside a specific set of deployment paths. For self-hosting, Moonshot recommends three inference engines: "Kimi K3 is recommended to run on the following inference engines" (raw.githubusercontent.com), naming vLLM, SGLang, and TokenSpeed, with quantization-aware training applied "using MXFP4 weights with MXFP8 activations for broad hardware compatibility" (raw.githubusercontent.com) specifically to widen the range of hardware able to serve the model, with the model weights and run recipes for each engine hosted on Hugging Face for self-hosted deployment.

This licensing structure matters for research and regulated-industry buyers specifically because it separates three cases with different obligations: purely internal use, use routed through Moonshot's own certified inference partners, and a third-party "Model as a Service" offering built on the weights. Section 2's separate-agreement rule applies only when the licensee or an affiliate operates a Model as a Service business and their aggregate revenue exceeds $20 million over any consecutive 12 months. Section 3's independent branding rule requires qualifying commercial products or services to display "Kimi K3" prominently when they exceed 100 million monthly active users or $20 million in monthly revenue. Sections 2 and 3 do not apply to internal use or use through Moonshot's official products or certified inference partners (github.com), so an organization self-hosting Kimi K3 for internal document analysis, rather than making the software, outputs, or underlying capabilities available to third parties, is outside both requirements.

Data Analysis and Evidence

Table 3 summarizes Kimi K3's own pricing structure across its two access paths: metered API access and flat-rate consumer membership.

Access method	Price	Notes	Source
API, input (cache miss)	$3.00 per 1M tokens	Flat rate, no length tiering	(www.kimi.ai) (platform.kimi.ai)
API, input (cache hit)	$0.30 per 1M tokens	A 90 percent reduction versus cache miss	(www.kimi.ai)
API, output	$15.00 per 1M tokens		(www.kimi.ai)
Kimi membership, entry tier	$19 per month (monthly billing)	Consumer chat access	(www.kimi.ai)
Kimi membership, top tier	$199 per month (monthly billing)	Consumer chat access	(www.kimi.ai)

The spread between Kimi K3's cache-miss and cache-hit input pricing, a 90 percent discount for repeated context, is the single largest lever in its cost structure for evidence-retrieval workloads that repeatedly query the same long document, since a cached multi-hundred-thousand-token source document is billed at roughly a tenth of its first-pass rate on every subsequent query.

Adoption signals collected directly from GitHub and Hugging Face on September 5, 2026 place the moonshotai/Kimi-K3 GitHub repository's engagement counts and the Hugging Face model page's like and follower counts in the low five figures each (api.github.com)(huggingface.co), though these are engagement counts rather than usage or revenue figures and should be read only as directional adoption signals. Separately, business press reported that Moonshot AI, the creator of Kimi K3, has filed for a Hong Kong initial public offering, with the South China Morning Post describing K3 at launch as "the world's largest open-weight AI model with more than 2.8 trillion parameters" (www.scmp.com). Reuters coverage relayed by Silicon Republic reported Moonshot was in early-stage talks with major cloud providers to host Kimi K3, "seeking up to a 30pc share of any revenue generated by Kimi K3-related services" (www.siliconrepublic.com), and separately that Moonshot's own comparative claims placed K3 as trailing "only OpenAI's GPT-5.6 Sol and Anthropic's Fable 5" (www.siliconrepublic.com) among named frontier competitors, a self-assessment broadly consistent with the independently measured Artificial Analysis Intelligence Index standing described above (artificialanalysis.ai).

Case Studies and Real-World Examples

GitHub Copilot: A Documented Third-Party Integration

On August 6, 2026, GitHub announced that "Kimi K3, an open-weight model, is now generally available in GitHub Copilot" (github.blog), across Visual Studio Code, Visual Studio, JetBrains IDEs, Xcode, Eclipse, the Copilot CLI, GitHub's mobile app, and github.com, with the model hosted through a third-party inference provider. This is a documented, neutral instance of a major software development platform adding K3 as a selectable model, independent of Moonshot's own marketing. GitHub's rollout was conditional rather than blanket: the changelog states "Kimi K3 is off by default for Copilot Business and Copilot Enterprise" (github.blog) customers, requiring an administrator to opt in through a policy setting before developers on paid organizational plans can select it, while individual Copilot users could enable it directly. The same changelog notes GitHub "temporarily paused the roll-out of Kimi K3 while [it mitigated] an incident with GitHub Actions" (github.blog), an availability issue on GitHub's own infrastructure that was unrelated to the Kimi K3 model itself, before resuming the rollout to general availability.

Community reception was not uniformly positive on cost. A Hacker News discussion thread titled "Kimi K3 is not cheap" (news.ycombinator.com) debated whether real-world token consumption matched the model's headline API pricing, a discussion consistent with Artificial Analysis's separate finding that Kimi K3's measured per-task cost is "particularly expensive when comparing to other open weight models of similar size" (artificialanalysis.ai). This is reported here as documented community sentiment, not as an independently verified technical finding.

Implications and Future Directions

For research-intensive fields such as life sciences, where a single evidence-retrieval task might mean reconciling clinical trial results, regulatory guidance, and internal study reports spanning hundreds of thousands of words, the practical question this report raises is not which vendor publishes the largest context-window number, but which

evaluation a reader can actually verify. IntuitionLabs, a life sciences and AI consultancy, frames this dynamic around a broader industry statistic on its own site: AI-enabled drug discovery and development work "can accelerate timelines by up to 60%," per Deloitte research the firm cites (intuitionlabs.ai), and a separate McKinsey estimate that "AI could generate $100B+ in annual value for the pharmaceutical industry" (intuitionlabs.ai) underscores why document-heavy, evidence-dependent workflows are a priority use case for long-context models generally, independent of any single vendor. Against that backdrop, this report's findings argue for evaluating a long-context model against the specific retrieval task at hand (agentic browsing, multi-document reasoning, or exact-fact recall at depth) rather than against window size alone, since Kimi K3's own data shows a full, uncompacted 1 million token context scoring slightly below a compacted 300,000 token strategy on the one retrieval-style benchmark where Moonshot published both figures (arxiv.org)(www.kimi.ai).

On "which model is best for long context evidence retrieval," the honest answer from currently published data is that no vendor, including Moonshot, has published a standardized needle in a haystack or RULER score for its newest flagship model that would let a reader compare exact-fact retrieval accuracy at depth across Kimi K3, GPT-6 Astra, Claude Opus 5, Gemini 3.1 Pro, DeepSeek V4, and Qwen3.8-Max on equal terms. Only the narrower AA-LCR reasoning benchmark in Table 1 and OpenAI's own MRCR disclosures for its own models (openai.com) (openai.com) offer independently sourced comparison points, and the two measure different capabilities. Given how closely the major vendors have converged on a roughly 1 million token window in 2026 (developers.openai.com) (platform.claude.com)(ai.google.dev)(ai.google.dev)(api-docs.deepseek.com)(www.alibabacloud.com), window size itself is likely to keep receding as a differentiator, while methodology transparency, independent verification of the kind Artificial Analysis provides for Kimi K3, and price-per-task rather than price-per-token alone are more likely to determine which model a research or regulatory-affairs team adopts for evidence-retrieval work over the next product cycle. Moonshot's stated pursuit of additional commercial distribution deals with major cloud providers (www.siliconrepublic.com) suggests broader enterprise availability of Kimi K3, potentially including regulated environments, is plausible in the near term, though no such deal had been finalized as of the September 2026 observation date.

Conclusion

Kimi K3 is a shipping, open-weight model independently confirmed to carry a 1,048,576 token context window (artificialanalysis.ai), built on a hybrid Kimi Delta Attention and Attention Residuals architecture that Moonshot describes in a public technical report (arxiv.org)(arxiv.org). On the one independently sourced long context reasoning benchmark in Moonshot's own disclosures, AA-LCR, K3's reported score of 74.7 edges out every comparison model Moonshot selected, including GPT-5.5 and Claude Fable 5 (github.com). On agentic evidence gathering (BrowseComp), K3 scores in the 90 to 91 percent range depending on context-management strategy (arxiv.org)(www.kimi.ai), and independent Artificial Analysis testing separately places it near the top of several agentic and knowledge-work benchmarks, while also noting a higher-than-typical per-token API cost relative to open-weight peers (artificialanalysis.ai). What is not available, from Moonshot or from any independent evaluator identified during this research, is a standardized needle in a haystack, RULER, or LongBench score for Kimi K3, which means the specific evidence-retrieval-at-depth claim implied by the model's headline context window remains, as of September 2026, unverified against the industry's original long context stress test. Readers evaluating Kimi K3 for document-heavy, evidence-dependent work should treat its 1 million token window as a

confirmed engineering fact, its AA-LCR and BrowseComp scores as the best currently available reasoning and retrieval-adjacent evidence, and the absence of a NIAH-style benchmark as an open question rather than a settled negative, pending future publication by Moonshot or a third party.

Frequently Asked Questions (FAQs)

What is Kimi K3's context window size? 1,048,576 tokens, commonly rounded to 1 million, confirmed independently by Artificial Analysis (artificialanalysis.ai) and in Moonshot's own technical documentation (arxiv.org).

How does Kimi K3 compare to GPT-4 in long context? It does not, in any current, verifiable sense: GPT-4 is a 2023-era OpenAI model that OpenAI no longer publishes pricing or benchmark data against, and the relevant current comparison is Kimi K3 against OpenAI's 2026 flagship, GPT-6 Astra, whose own published window is 1,050,000 tokens (developers.openai.com).

Does Kimi K3 have published needle in a haystack results? No standardized needle in a haystack, RULER, or LongBench score for Kimi K3 was located in Moonshot's own materials or among independent evaluators as of the September 2026 observation date; Moonshot instead publishes BrowseComp and AA-LCR figures (arxiv.org) (github.com).

What does the Kimi K3 API cost? $3.00 per million input tokens on a cache miss, $0.30 per million on a cache hit, and $15.00 per million output tokens (artificialanalysis.ai).

Is Kimi K3 open source? It is open-weight under a custom Kimi K3 License with commercial-scale conditions, not a standard open-source license (raw.githubusercontent.com)(huggingface.co).

Which independent organization benchmarks Kimi K3? Artificial Analysis, which reports an Intelligence Index score of 57 for Kimi K3 (artificialanalysis.ai).
