# Episode Book

EPISODE #6253

# Why Context Windows Stalled at a Million Tokens

| | |
|---|---|
| PUBLISHED | 24 October 2026 at 15:07 UTC |
| DURATION | 19:14 |
| LISTEN | myweirdprompts.com/episode/context-window-million-token-limit-38791/ |
| HOSTS | Corn, Daniel, Herman, Hilbert |
| SCRIPT MODEL | anthropic/claude-sonnet-5.5 |
| VOICES | chatterbox-regular+meorav (chatterbox-local) |
| PIPELINE | V5.22 |
| BOOK GENERATED | 25 October 2026 at 06:19 UTC |

## SYNOPSIS

Language model context windows grew thirtyfold a year before hitting a ceiling at one million tokens. Hardware limits explain why: storing and constantly reading massive key-value caches rapidly outgrows flagship GPU memory and bandwidth. Compounding the cost, real-world retrieval accuracy plummets across long documents, leaving labs to double prices for lengthy queries while pushing context compaction over brute-force expansions.

# TRANSCRIPT



**Corn**

Two million tokens. That's the window xAI was selling on its fast Grok models at the end of last year. The newest Grok models list five hundred thousand.



**Daniel**

That's the number that stopped me. Everything else about these models moves every few months, the benchmarks, the prices, the speed, and the context window has sat at about a million tokens for a year. Anthropic, OpenAI, Google, all parked in the same spot. So what's actually holding it there, and what would break it? Because I keep seeing people claim ten million, a hundred million, and I don't know who to believe.



**Corn**

It hasn't only stalled, either. Google's two million tier from Gemini one point five Pro is retired. Gemini two point five Pro was advertised with two million coming soon, and it shipped at one. In a couple of places the ceiling went sideways, and in one place down.



**Herman**

Which is odd against the history. Epoch tracked the longest windows growing around thirty times a year from the middle of twenty twenty-three. Gemini one point five hit a million in February twenty twenty-four. It's now October twenty twenty-six. Thirty times a year for two and a half years would put us somewhere ridiculous, and instead there's a flat line.



**Corn**

A trend that quits that completely hit a wall. It didn't lose interest.



**Daniel**

So what's the wall? Is it that nobody knows how to build a longer one?



### Herman

No, they can build it. Meta put out Llama four Scout with a ten million claim, and I'll get to why I wouldn't lean on that. The wall is cost, and it has a very particular shape. Memory first. When a model reads your prompt, it stores a key and a value for every token at every layer, and that stash is the KV cache. Take a seventy-billion-parameter dense model of the Llama three point one class. Eighty layers, eight key-value heads, sixteen-bit numbers. That works out to roughly three hundred twenty kilobytes for every single token.



### Daniel

That doesn't sound like much.



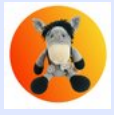
### Herman

Per token, no. Multiply by a million, because Google's window is one million tokens on the nose. About three hundred twenty-eight gigabytes, for one user's conversation. The best GPU you can buy right now, the B three hundred, has two hundred eighty-eight gigabytes of memory in total.



### Corn

One person's chat history outgrows the entire chip. You'd have to ask the GPU to move over.



### Herman

It fills about half of an eight-GPU H one hundred node, so you're renting a small rack for one person's afternoon. And that's before the second problem. Every token the model writes means re-reading that whole cache. Capacity has grown on the newer chips, but the B three hundred's memory bandwidth is eight terabytes a second, the same as the B two hundred. So a long session spends its life waiting on memory, and the number of people you can serve at once collapses.



### Daniel

Hang on, is that the same problem as the quadratic thing? The attention cost going up with the square of the length?



### Herman

Related, different bill. The quadratic part hits when the model first reads the prompt, the prefill. Ten times the tokens is roughly a hundred times the attention compute, so going from one million to ten million is a hundredfold jump. The memory problem hits every token after that. You pay one at the start and the other forever.



### Corn

Hundredfold sounds like the sort of number a lab quotes in a press release, not in a budget meeting.



### Herman

It's the number in the budget meeting, believe me. But I'm making the dense-cache story sound worse than it is now, because the newest models don't store a dense cache like that. DeepSeek's V four Pro, at a million tokens, needs about a tenth of the cache its predecessor did, and the Flash model gets down to seven percent. Against a standard grouped-query cache it's about two percent.



### Daniel

So if the cache problem is being solved, why isn't everyone at two million?



### Herman

Nobody has said. I haven't seen a lab explain why it's one million and not two, so anything I tell you is reading tea leaves. What I can read is the pricing, because the pricing shows the labs' real cost curve. OpenAI's GPT six Astra takes about a million tokens, but once a request goes past two hundred seventy-two thousand input tokens, the whole request costs double on input and one and a half times on output. The whole request, not just the part past the line.



### Corn

The window is a million tokens and the price tag starts looking embarrassed at a quarter of that.



### Herman

Google does the same on Gemini three point one Pro, from two dollars to four dollars per million input tokens above two hundred thousand. xAI doubles at two hundred thousand too. Anthropic is the odd one out. The full million at one flat price, with Haiku five point five as the exception, which charges more past a hundred thousand.



### Daniel

Why would Anthropic eat that cost?



### Herman

I don't know, and I'd be guessing. They may simply have the serving efficiency. Or it's a bet that most people send short requests anyway and the long ones are the exception. Corn, you'd know more about betting that people stay lazy than I would.



### Corn

That one's on the record as a hobby. We've got someone on the line with the plainest version of this question, so let's take it.
Bruce: Sorry, I've been nodding along for months and never asked. What actually is a token, is it like a word, or more like a letter, and what's this context window, is it the box on the screen where you type?



### Herman

Not at all. A token is a chunk of text, usually a bit shorter than a word, and the context window isn't the box you type in, it's how much the model can keep on the table in front of it at once, like a kitchen table where, once it's full, things slide off the edge... I did just spend ten minutes on memory bandwidth, and I'd like that noted.



### Corn

Noted, and the table is full. Herman, the other half of this is whether anybody wants more table. The window a lab advertises and the window a model can actually use are different numbers.



### Herman

DeepSeek published a chart that shows it. On an eight-needle retrieval test, V four scores above point eight two out to two hundred fifty-six thousand tokens, and at a million it's down to point five nine. That's a lab's own number on its own model, and it still loses about a third of its accuracy by the end. Epoch's estimate for Gemini two point five Pro was that it was effectively good to about eight hundred seventy-seven thousand of its million.



### Daniel

So a million tokens is a headline and not a guarantee.



### Herman

It's a headline with a gradient. Epoch also measured the length where top models stay at eighty percent accuracy, and that grew more than two hundred fifty times in nine months, up to mid twenty twenty-five. So the usable length was catching up with the advertised length. Then the advertised length stopped, and the usable one is still catching up. A compilation of one-million-ish retrieval scores, older models mind you, had Opus four point six around seventy-six percent, GPT five point four at thirty-seven, and Gemini three Pro at about twenty-five.



### Corn

Same window, a threefold spread in how much of it works. That's a bigger gap than the one between two and one.



### Daniel

Is that the lost-in-the-middle problem again, or something else?



### Herman

Same family. The model attends to the start and the end better than the middle, and the longer the prompt, the more middle there is. My guess on why the labs stop at a million is partly that. Past that point you'd mostly be selling space the model can't read well. And training at long lengths is expensive, which is why Scout trained short and stretched. Please don't quote me on that, because no lab has said it.



**Corn**

Anthropic's own documentation more or less says so. It tells you more context isn't automatically better, that recall degrades, and that for long-running agents the main strategy is compacting the history, not buying a bigger window.



**Herman**

Which tells you where the product pull is. If the agent summarises as it goes and retrieves what it needs, nobody is begging for a hundred million tokens. I think that's the least discussed reason the ceiling held. The labs can't build it cheaply, and the customers aren't shouting loudly enough for them to try.



**Corn**

Somebody is always shouting for more. I'd say that's most of the internet.



**Herman**

Epoch makes the loud case. Their arithmetic is that a million tokens holds about thirty minutes of an agent working on a computer, and ten million holds around six hours. Forty million might hold months of what you'd call work experience. So tell me, Corn, you'd rather a model remembered everything. Would you pay double for it, if it worked?



**Corn**

If it worked, yes. The "if" is carrying a lot of weight in that sentence. I'd pay double for ten million and then spend the money finding out whether the model read the middle.



**Daniel**

Okay, so what are the claims past a million? Because some of those headlines are loud.



**Herman**

There are three families. The first is retrieval claims. Subquadratic launched in May with a twelve-million-token API, and a third-party test reported ninety-eight percent on needle-in-a-haystack at both six and twelve million. We spent a whole episode on it, When



### Herman (continued)

a Startup Claims to Break the Quadratic Wall. What matters is what's missing. No independent reasoning result at twelve million, access is a waitlist, and it's reportedly built on reused Qwen weights.



### Corn

Ninety-eight percent at finding one sentence. A search box does that.



### Herman

Magic claimed a hundred million tokens in August of twenty twenty-four and raised around three hundred twenty million dollars. No API, no paper, no product since. The one thing it left behind is a benchmark.



### Daniel

And Llama Scout's ten million?



### Herman

Trained at two hundred fifty-six thousand tokens, then stretched with tricks to generalise longer. MiniMax did something similar years ago, trained on a million and stretched to four. I haven't found an independent evaluation of either at their top end, so the numbers stand on a claim.



### Corn

Three claims and zero people who've checked the middle.



### Daniel

Is there anything past a million that someone has actually tested for reasoning?



### Herman

Thinly. A benchmark called AgentLongBench, built around long agent tasks, reports that Grok four point one stays above fifty out to two million tokens, while GPT four point one and Gemini two point five Flash decay after two hundred fifty-six thousand, and open-weight models fall to negligible by a million. I'm going from the abstract, so hold that loosely. It's the only reasoning-style result past a million that I know of.



### Corn

Which means the one model that tested well is also the one whose window just got cut to a quarter.



### Herman

I noticed that too. I can't tell you why xAI did it. Lower window, lower serving bill is my guess. The other result worth knowing is not a bigger window at all. Recursive language models, from MIT. The long prompt sits outside the model as a variable in a code environment, and the model writes code to slice it and calls itself on the slices. The authors report working past ten million tokens on a browsing benchmark.



### Daniel

That sounds like cheating.



### Herman

A little. An independent reproduction found real gains on hard reasoning, but simple retrieval got worse, one hundred percent down to eighty-five at one setting, and it was much slower.



### Corn

So it can read ten million tokens, and it takes a long weekend.



**Daniel**

What about changing the attention itself? Is that the other route?



**Herman**

That's where the engineering is moving, and it's messy. MiniMax tried linear attention, hit quality problems, and went back to sparse and full attention in its newer models. Meanwhile Qwen and NVIDIA's Nemotron run hybrids, mostly cheap layers with a few real attention layers mixed in, and both land at a million. Qwen's open weights are natively two hundred sixty-two thousand and only reach a million with an extension trick. Nobody has dropped real attention completely and kept reasoning quality.



**Corn**

Every shortcut has a toll booth on the far end.



**Daniel**

What does a hybrid actually give up, though?



**Herman**

Exact recall, as I understand it. A full attention layer can look back at any earlier token directly. The cheap layers keep a fixed-size running summary instead, so the cache stops growing, but a detail from nine hundred thousand tokens ago can get blurred. That's why the hybrids keep a few real attention layers in the stack. NVIDIA's Nano two paper replaced most of the attention with Mamba and reports three to six times the throughput.



**Corn**

You've been waiting all episode to say three to six times, haven't you.



**Herman**

Since the start of the memory section, yes. Don't take it away from me.



**Daniel**

And the output side? I saw Google doing something there.



**Herman**

Gemini four Argon lists the same one million forty-eight thousand token window, but it lets the output run up to a million tokens with a mode they call long decode, against sixty-five thousand by default. That's the new thing, a model writing a book-length answer. It's still the cheap end of the problem, because the model generates it one token at a time and nobody has to read in a million all at once.



**Daniel**

So forecasts. Does anyone put a number on the ceiling moving?



**Herman**

A prediction market has about a thirty-nine percent chance of any public frontier model with ten million or more by the end of next year. Take that the way you take prediction markets, but it isn't zero. Epoch's view is that cost barriers don't look insurmountable. NVIDIA announced a prefill-focused chip, Rubin CPX, aimed at million-token work, though I've seen a claim that it got cancelled, and I can't confirm that either way.



**Daniel**

What would you bet on, Corn? What breaks it?



**Corn**

Not a bigger window. A cheaper one. If a one-million-token request costs the same as a short one everywhere, and Anthropic already looks like that, then ten million stops being a cost problem and becomes a quality problem, which is the one nobody has solved. I'd watch for the first model where the retrieval curve doesn't sag at the end.



**Herman**

I'd split it. The window grows when the hybrid architectures prove they hold quality, and in the meantime agents never need the whole thing in one go. My worry for the big-window crowd is that the workaround, compaction and retrieval, is good enough that nobody funds the hard fix.



**Daniel**

Would a lab just publish a checkable long-context reasoning number, though?



**Herman**

They could. I'd call it the clearest sign. A lab with a ten-million window that really reasons will publish the curve before it publishes the price.



**Corn**

You've turned into a prediction market, Herman. Pick a number.



**Herman**

Before the end of twenty twenty-seven, I'd say one lab ships a window past a million, probably with a hybrid architecture. I wouldn't put it past two million.



**Corn**

That's a donkey hedging. I'm writing it down.



**Herman**

Fine. Here's one for you. If somebody handed you ten million tokens that worked, what would you actually put in it?



**Corn**

Everything about something I distrust, all at once, and then ask what it's hiding. A whole company's filings, every email in a lawsuit. I don't want to remember more. I want to be able to check more.



**Daniel**

Does any of this matter to somebody who just types questions into a chat?



**Herman**

Mostly not, my guess is. Very few ordinary conversations get anywhere near two hundred thousand tokens, so the price cliffs never bite. It matters for agents that run for hours, read whole code bases, and drag their own history behind them. That's where a flat price or a bigger window starts to show up on a bill.



**Corn**

So the ceiling is a problem for the people building the thing, and a rumour for everyone else.



**Hilbert**

Brother P-touch Cube. Thirty-nine dollars. I own thirty-one of them, one for each assistant I talk to, and every answer gets printed on tape as it comes. Aunt Berta hangs the strips down the stairwell. When the tape reaches the third floor, that conversation is finished, and I don't care what the model says its limit is.



**Herman**

Thermal tape pales over time, so you'd want those strips out of the sunlight.



**Hilbert**

She keeps the stairwell shut. Herman, you went round the houses on the memory arithmetic. Also, it isn't a million on the nose. It's one million forty-eight thousand five hundred seventy-six. Two to the twentieth. A cassette of tape is six dollars and Berta gets



**Hilbert (continued)**

exactly one conversation out of it.



**Corn**

Berta will have to rehang the whole lot when you move flat. That's your real cap.



**Hilbert**

We're not moving. She'll want a hook every thirty centimetres, though. I'll get hooks.



**Herman**

One more from the history that didn't fit. Google DeepMind reportedly had a ten-million-token version of Gemini one point five Pro and declined to release it because of cost. That's Epoch's wording, reportedly, so I can't tell you it's confirmed.



**Corn**

So the ceiling may have been a pricing decision from the start.



**Daniel**

Which would mean the limit isn't that nobody can build it. Nobody can serve it at a price, and nobody's shown it thinks well at the far end.



**Herman**

Keeping a long conversation costs somebody, whether it's tape at six dollars a cassette or memory on a chip. Whoever makes that cheap is the one who moves the ceiling.



**Corn**

Thanks to our producer Hilbert Flumingtop for keeping the whole thing running. This has been My Weird Prompts.



**Herman**

Every episode is at my weird prompts dot com.



**Daniel**

Cheers, both of you. See you tomorrow.



**Corn**

See you tomorrow.