MY WEIRD PROMPTS

Podcast Transcript

EPISODE #111

Beyond Transformers: Solving the AI Memory Crisis

Published December 27, 2025 • Runtime: 21:46

https://myweirdprompts.com/episode/ai-stateless-architecture-future/

EPISODE SYNOPSIS

In this episode, Herman and Corn Poppleberry tackle one of the most frustrating hurdles in modern AI engineering: the "stateless" architecture of Large Language Models. They explore why current models require you to resend your entire conversation history with every message, leading to skyrocketing token costs and the "lost in the middle" phenomenon that plagues even the most advanced systems. From the quadratic complexity of the standard Transformer to the revolutionary potential of State Space Models like Mamba and hybrid architectures like Jamba, the brothers break down how researchers are finally building AI with persistent, human-like memory.

DANIEL'S PROMPT

Daniel

Certainly! Here's a cleaned transcript of your prompt: "We've discussed how context is vital for getting reliable and performant results from AI, yet many tools struggle with context pruning or limiting the context trail. Large language model APIs typically use a stateless architecture, meaning each new turn in a conversation requires resending the entire previous history. This leads to context aggregation and significant API costs. Given these challenges, why is a stateless architecture the default for LLMs? Are there any fundamental architectural proposals beyond the Transformer model that could make AI better suited for conversational use without these limitations?"

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts! I'm Corn, and I am joined as always by my brother.

Herman

Herman Poppleberry, reporting for duty! It is December twenty-seventh, two thousand twenty-five, and we are closing out the year with a real brain-tickler.

Corn

Yeah, we are. So, our housemate Daniel sent us a prompt today that actually makes me feel a little better about my own memory. You know, as a sloth, people expect me to be a bit... slow on the uptake, but it turns out even the most advanced AI has a bit of a memory problem.

Herman

It really does. Daniel was asking about context and why these massive language models seem to struggle with it. Specifically, he’s curious about stateless architecture. Why is it the default, and are there better ways to build these things so they don’t have to reread the entire conversation every time we hit enter?

Corn

Right, because Daniel mentioned that every time he talks to an AI through an API, it’s like the AI has a total reset. It doesn’t remember the last thing he said unless he sends the whole history back to it. That sounds... exhausting. And expensive!

Herman

It is both of those things, Corn. It’s actually one of the biggest hurdles in AI engineering right now. We’re in late two thousand twenty-five, and while we’ve seen models with massive context windows—some handling millions of tokens—the underlying way they process that information is still surprisingly clunky.

Corn

Okay, so let’s start with the basics for people like me. What does "stateless" actually mean in this context?

Herman

Think of it like a very polite, very smart waiter who has absolutely no short-term memory. You sit down and say, I’d like a coffee. The waiter goes to the kitchen, brings you a coffee, and then immediately forgets you exist. When you want a refill, you can’t just say, another one please. You have to say, hi, I am the person who sat down two minutes ago, I ordered a black coffee, it was delicious, and now I would like a second black coffee.

Corn

That sounds like a terrible way to run a restaurant. The waiter would be exhausted, and I’d be annoyed.

Herman

Exactly! But in the world of Large Language Models, or LLMs, that’s how the APIs work. Every request is a fresh start. The model itself doesn’t "hold" your conversation in its brain between turns. To get it to understand a conversation, you have to bundle up every single previous message and send it back to the server every single time.

Corn

So if I’ve been chatting for an hour, the "packet" of info I’m sending gets bigger and bigger?

Herman

Precisely. This is called context aggregation. And since these companies charge you by the token—which is basically a word or a piece of a word—your twenty-first message costs way more than your first message, because you’re paying to send all twenty previous messages again.

Corn

That seems like a design flaw. Why would the smartest people in the world build it that way? Is there a reason it has to be stateless?

Herman

It’s not so much a flaw as it is a trade-off for scale. Imagine you’re a company like OpenAI or Anthropic. You have millions of people talking to your AI at the same time. If the AI had to "remember" every single conversation in its active memory—meaning, if it was stateful—the server requirements would be astronomical.

Corn

Oh, I see. So by being stateless, the server can just handle a request, finish it, and immediately move on to the next person without having to keep a "folder" open for me?

Herman

Spot on. It makes the system much easier to scale and load-balance. You can send my first message to a server in Iowa and my second message to a server in Belgium, and it doesn’t matter because I’m sending all the context anyway. If it were stateful, I’d have to stay connected to that one specific server that "remembers" me.

Corn

Okay, that makes sense from a business perspective, but it’s a bummer for the user. Daniel mentioned "context pruning" and "context muddling." What’s happening there?

Herman

Well, as that "packet" of history gets longer, two things happen. First, it gets expensive. Second, the model starts to lose the thread. Even though we have models now that claim to have a context window of two million tokens, they still suffer from what researchers call the "lost in the middle" phenomenon. They remember the very beginning of the prompt and the very end, but they get a bit hazy on the details in the middle.

Corn

I get that. If I read a thousand-page book in one sitting, I might remember how it started and how it ended, but page five hundred forty-two might be a bit blurry.

Herman

Exactly. And "pruning" is the process of trying to cut out the fluff so you don't hit those limits or pay too much. But if you prune the wrong thing, the AI loses the context it needs to give a good answer. It's a delicate dance.

Corn

It sounds like we're trying to fix a leaky faucet by just putting a bigger bucket under it. Is there an actual architectural fix? Like, is the Transformer model itself the problem?

Herman

That is the million-dollar question, Corn! Or maybe the trillion-dollar question given the current AI market. The Transformer architecture, which is what powers almost every major LLM today, has a specific mathematical property called quadratic complexity.

Corn

Whoa, slow down. "Quadratic complexity"? Speak sloth to me, Herman.

Herman

Haha, sorry! Basically, it means that if you double the amount of text you want the AI to look at, the amount of computational work the AI has to do quadruples. If you triple the text, the work increases nine-fold. It's not a one-to-one increase. It gets exponentially harder for the model to "pay attention" to everything as the text gets longer.

Corn

That sounds like a recipe for a crash. No wonder it's so expensive.

Herman

It really is. And that’s why researchers are looking for something "beyond the Transformer." But before we get into the heavy-duty engineering stuff, I think we have a word from someone who definitely doesn't have a memory problem... or maybe he does.

Corn

Oh boy. Let's take a quick break for our sponsors. Larry: Are you tired of your brain feeling like a browser with too many tabs open? Do you walk into rooms and forget why you’re there? Introducing the Echo-Memory Three Thousand! It’s not a hearing aid, it’s a life-recorder! This sleek, slightly heavy lead-lined headset records every single word you say and every word said to you, then plays it back into your ears on a four-second delay! Never forget a grocery list again because you’ll be hearing yourself say "eggs" while you’re standing in the dairy aisle! Side effects may include mild vertigo, temporal displacement, and the inability to hold a conversation without weeping. The Echo-Memory Three Thousand—because the past is always louder than the present! Larry: BUY NOW!

Herman

...Thanks, Larry. I think I’d rather just forget the grocery list, honestly.

Corn

I don’t know, Herman. A four-second delay sounds like a great excuse for me to take even longer to answer questions.

Herman

You don’t need any help in that department, brother. Anyway, back to the serious stuff. Daniel asked if there are fundamental architectural proposals beyond the Transformer that could fix this stateless, context-heavy mess. And the answer is a resounding yes. We’re seeing a massive shift in research toward something called State Space Models, or SSMs.

Corn

State Space Models. Okay, how do those differ from the Transformers we’ve been using?

Herman

The big one people are talking about in two thousand twenty-five is called Mamba. It was originally proposed a couple of years ago, but it's really hitting its stride now. The magic of Mamba and other SSMs is that they have linear complexity.

Corn

Linear complexity. So, if I double the text, it only doubles the work?

Herman

Exactly! No more quadratic explosion. This means, theoretically, you could have a context window that is effectively infinite without the cost or the compute time blowing up.

Corn

That sounds like a game-changer. How does it actually do that? How does it "remember" without rereading everything?

Herman

It works more like a traditional Recurrent Neural Network, or RNN, but with a modern twist. Instead of looking at every single word in relation to every other word—which is what the "Self-Attention" mechanism in a Transformer does—an SSM maintains a hidden "state." This state is like a compressed summary of everything it has seen so far.

Corn

So it's like the AI is taking notes as it reads?

Herman

That's a perfect analogy! As it reads word one, it updates its notes. When it gets to word one hundred, it doesn't have to look back at word one; it just looks at its notes. This makes it much more like a human conversation. When I'm talking to you, I don't re-process every word you've said since nine A.M. I just have a "state" in my head of what we're talking about.

Corn

So why aren’t we all using Mamba right now? Why is everyone still obsessed with Transformers?

Herman

Well, Transformers are incredibly good at "recalling" specific facts from a huge pile of data. They’re like having a photographic memory of the whole page. SSMs are great at the "flow" and the "summary," but they can sometimes struggle with very precise retrieval—like if you asked it for the third word on page seventy-two of a massive document.

Corn

Ah, the notes aren’t as good as the original text.

Herman

Exactly. But here’s the cool part: in two thousand twenty-five, we’re seeing "hybrid" models. There’s a model architecture called Jamba, for instance, that mixes Transformer layers with Mamba layers. It tries to give you the best of both worlds—the precision of a Transformer and the efficiency and "memory" of a State Space Model.

Corn

That’s fascinating. It’s like having a guy who takes great notes but also has a few pages of the original book memorized just in case.

Herman

Precisely. And there’s another approach called "Retentive Networks" or RetNet. They claim to have the parallel training of Transformers but the efficient inference of RNNs. Basically, they want to be fast when you’re training them and fast when you’re talking to them.

Corn

So, does this mean the "stateless" problem goes away? Will Daniel finally be able to talk to an AI without sending his whole life story back every time?

Herman

We're getting there. Some of these newer architectures allow for "stateful" APIs. Instead of you sending the history, the server just keeps that "compressed note" or "state" active for your session. Because the state is so much smaller than the full text history, it's actually feasible for the company to store it for you.

Corn

That would save so much money on tokens.

Herman

Oh, absolutely. It would slash API costs. And it would make the AI feel much more like a persistent companion. You wouldn't have to "remind" it of who you are or what your project is every time you open the chat. It would just... know.

Corn

That sounds a bit like the "Personal AI" dream people have been talking about for years.

Herman

It is! And we're also seeing progress in something called "KV Caching," which stands for Key-Value caching. It's a way for current Transformers to be a little less forgetful. It essentially saves the mathematical "work" the AI did on the previous parts of the conversation so it doesn't have to re-calculate everything from scratch.

Corn

Wait, so if they can already do that, why is it still expensive?

Herman

Because even if you cache the "work," you still have to load all that data into the high-speed memory of the GPU—the graphics chip—every time you want to generate a new word. And GPUs have very limited high-speed memory. It’s like having a tiny desk. You can be the fastest worker in the world, but if your desk is only big enough for three papers, you’re going to spend all your time swapping papers in and out of your drawers.

Corn

I feel that. My "desk" is basically just a pillow, and it’s always full.

Herman

Haha, exactly. The "desk" is the VRAM—the Video RAM—on the chip. These new architectures like Mamba or hybrid models are designed to use a much smaller desk. They’re more efficient with their "workspace," which means they can handle much longer conversations without slowing down or costing a fortune.

Corn

So, looking forward to two thousand twenty-six, do you think we’re going to see a "post-Transformer" world?

Herman

I think "hybrid" is the keyword for next year. We’re going to see models that use Transformers for the heavy lifting and deep reasoning, but use these State Space layers for the "memory" and the long-term context. It’ll make AI feel less like a series of disconnected prompts and more like a continuous stream of thought.

Corn

That’s actually a bit reassuring. It makes the AI feel a little more... well, human. Or at least, a little more like a donkey with a very organized filing cabinet.

Herman

Hey, I’ll take it!

Corn

So, for the regular people listening, or for Daniel when he’s working on his projects, what’s the practical takeaway here? Is there anything we can do right now to deal with this stateless mess?

Herman

Well, until these hybrid models become the industry standard, the best thing you can do is "context management." First, use "system prompts" effectively. Instead of putting all your instructions in every message, put them in the system prompt—some APIs cache that specifically to save you money.

Corn

Okay, system prompts. What else?

Herman

Second, be your own "pruner." If a conversation gets too long and the AI starts acting weird or "muddled," start a new session but give it a concise summary of the important points from the last session. You’re basically doing the work of that "hidden state" we talked about.

Corn

Like a "Previously on My Weird Prompts" recap.

Herman

Exactly! And third, keep an eye on the smaller, specialized models. Sometimes a smaller model with a specialized architecture for long context—like some of the newer "long-context" versions of Llama or Mistral—will actually perform better and cheaper for a long conversation than a massive, "general" model that’s struggling under its own weight.

Corn

That’s a great point. Bigger isn’t always better, especially if the bigger model is paying a "quadratic tax" on every word.

Herman

Precisely. We’re moving from the era of "just add more layers" to the era of "make the layers smarter." It’s a shift from brute force to elegant engineering.

Corn

I like that. It sounds much more sustainable. And hopefully, it means Daniel won't have to spend his whole rent on API tokens just to get his AI housemate to remember where he left his keys.

Herman

Haha, well, let’s hope the AI is better at finding keys than we are, Corn. We still haven't found that spare set for the balcony.

Corn

Don't look at me, I'm still trying to remember what I had for breakfast.

Herman

It was a leaf, Corn. It’s always a leaf.

Corn

Ah, right. Good state management, Herman.

Herman

I try!

Corn

Well, this has been a fascinating deep dive. It’s amazing how much of the "intelligence" we see in AI is actually held back by these basic architectural plumbing issues.

Herman

It really is. It’s like having a genius stuck behind a very slow dial-up connection. Once we fix the "connection"—the way the model handles state and context—we’re going to see another massive jump in what these things can actually do in our daily lives.

Corn

I’m looking forward to it. Thanks for breaking that down, Herman. You made "quadratic complexity" sound almost... simple.

Herman

My pleasure, brother. It’s what I’m here for.

Corn

And thanks to Daniel for the prompt! If you’re listening and you’ve got a weird question or a topic you want us to dig into, head over to myweirdprompts.com and use the contact form. We love hearing from you.

Herman

You can also find us on Spotify and anywhere else you get your podcasts. We’ve got a whole archive of us trying to make sense of this wild world.

Corn

This has been My Weird Prompts. We'll see you in the new year, everyone!

Herman

Happy New Year!

Corn

Bye! Larry: BUY NOW!