Episode Book

EPISODE #6211

Attention Is Still Winning: The Hybrid Model Census

PUBLISHED	22 October 2026 at 22:44 UTC
DURATION	21:04
LISTEN	myweirdprompts.com/episode/hybrid-attention-model-census-36536/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
BOOK GENERATED	23 October 2026 at 07:44 UTC

SYNOPSIS

Nine months after our quadratic bottleneck episode, the post-transformer world has actually shipped — just not in pure form. This episode is a census of the hybrid models now in the wild: Qwen 3.5's three-to-one Gated DeltaNet blocks, Nvidia's Nemotron 3, Kimi Linear's 75% cache cut, IBM's Granite 4, Ant Group's trillion-parameter Ling and Ring, and Falcon H1's side-by-side design. Then the counter-story: MiniMax built a linear-attention model, went back to full attention, and published exactly why — expensive evaluation, immature infrastructure, and multi-hop reasoning that broke at scale. Plus the tradeoff underneath it all: fixed-size state means fixed-size memory, and a phrase from token 40 can go missing at token 500,000. What survives is a division of labour — a few attention layers acting as the lookup desk for many cheap recurrent ones.

TRANSCRIPT



Corn

Is anything you actually chat with today not a transformer? That's the thing Daniel wants settled. Back in January we did the quadratic bottleneck episode, Mamba and state space models and why attention gets ruinously expensive as the input grows, and at the time all of it felt like a research toy. Nine months on, he wants to know whether the post-transformer world has shipped, and what's happening to attention and the KV cache underneath.



Herman

It has shipped, just not in the shape January implied. Mamba-style layers sit inside real, downloadable, benchmarked models now. Almost none of those models are pure, though. They're hybrids, a lot of cheap non-attention layers with a few attention layers kept in. And the non-attention layer you meet most in the big open models isn't Mamba. It's a cousin.



Corn

So the revolution arrived with a spare transformer in the boot, just in case.



Herman

Pretty much. Take Qwen three point five, the big open model from February. Sixty layers, laid out as fifteen identical blocks. Each block is three Gated DeltaNet layers, then one ordinary gated attention layer. Three to one. Three hundred and ninety-seven billion parameters in total, about seventeen billion active per token.



Corn

Before the census, one thing we're not redoing. The hybrid idea isn't new to the show. We did the AI twenty-one Jamba episode in March, and the attention-variants episode on multi-query, grouped-query and latent attention, and the recent KV cache ones. Those were all about making the transformer cheaper from inside. This is about what happens when the layers themselves change.



Herman

Right, and the two stories are now tangled together, because the same models do both. Compressed attention in one layer, no attention at all in the next three. Hold that thought.



Corn

And Gated DeltaNet is not Mamba.



Herman

Same family. A linear recurrent layer with a fixed-size state, so it reads the text once, left to right, and carries a summary forward. The difference is in how the summary is updated. Mamba decays the old and adds the new. The delta rule is more like editing a notebook: the layer checks what its memory currently predicts for a key, and corrects it toward the new value. A gate on top lets it wipe things fast. In the comparisons people ran, that correction step recalls better than plain decay.



Corn

Better than Mamba. Then why does anybody still say Mamba?



Herman

Because Nvidia didn't switch. The Nemotron three family is Mamba two layers, mixture of experts, and a handful of attention layers. The small one has twenty-three Mamba layers and six attention layers in a stack of fifty-two. The Super, the hundred-and-twenty-billion one from March, is closer to five Mamba layers per attention layer. And even its attention is thin, only two key-value heads.



Corn

Six attention layers out of fifty-two. That's less transformer than my average afternoon has wakefulness.



Herman

Ha. Fine, but the cache is where it pays. Attention layers are the only ones that grow a KV cache with the length of the conversation. A Mamba or DeltaNet layer carries the same fixed-size state at a thousand tokens or a million. So if only one layer in four holds a cache, you've cut the cache by about three quarters before doing anything clever.



Corn

Which is roughly what Moonshot claimed with Kimi Linear.



Herman

Up to seventy-five percent less cache, and up to six times faster decoding at a million tokens. Their linear layer is a refined Gated DeltaNet they call Kimi Delta Attention, mixed three to one with DeepSeek-style compressed attention. There's talk of a much bigger Kimi K three at trillions of parameters with the same pattern, but I've only seen that in one secondary write-up, so I'd hold it loosely until the model card says it.



Corn

Holding it loosely.



Herman

The ratio isn't universal either. IBM's Granite four family is nine Mamba layers for every transformer layer, Apache licensed, aimed at cheap memory use. IBM says over seventy percent less RAM on long inputs and many parallel requests. Ant Group has a trillion-parameter Ling and Ring pair, hybrid linear with selected full layers. And Falcon H one out of Abu Dhabi does something else again: attention and Mamba run side by side inside every block, instead of in separate layers.



Corn

Side by side. Two people reading the same page and muttering.



Herman

Not far off. At the small end, Liquid AI's LFM two models use short gated convolutions plus a few attention layers, and they found that mix by searching for what runs fast on a phone processor. And Cartesia, whose founders co-created the structured state space line that Mamba grew from, runs state space models in a streaming speech product people really use. Their latency numbers are vendor claims and independent measurements I've seen are slower, but it's real.



Corn

That's a lot of names. Is the list getting longer or just louder?



Herman

Longer, I think, and the scale is the news. A year ago a hybrid meant a few billion parameters from a research group. Now it's a four-hundred-billion Qwen and a trillion-parameter Ant Group model, and Qwen's own follow-ups this spring are reported to use the same recipe. I'm going from search snippets on those, so treat the exact sizes as soft. A smaller thing worth saying: the AI twenty-one side of this has gone quiet. Jamba two appeared in January, and I've seen only a single reseller page claiming the company has since moved toward an orchestration product. I wouldn't lean on it.



Corn

So the architecture is doing better than the company that popularised it.



Herman

That happens a lot, doesn't it. The idea escapes the lab that had it.



Corn

Here's what I don't get. Three to one, five to one, nine to one. If the cheap layer is that good, why keep any attention at all? Why not go pure?



Herman

A fixed-size state is a fixed-size memory. It has to decide, while the text streams past, what's worth keeping. Attention decides nothing up front. It keeps everything and looks back when asked. If the question at token nine hundred thousand hinges on one odd phrase from token forty, the state may have squeezed it out. There's a twenty twenty-four result showing transformers beat these models at plain copying, and the authors are careful to say it applies to any fixed-memory model, not Mamba in particular.



Corn

Make it concrete for me. What does losing a phrase from token forty look like in practice?



Herman

Picture a long agent session. Forty thousand tokens in, the user says their staging server is called the name of some bird. Five hundred thousand tokens later, a tool output says deploy failed on the bird. A model with exact lookup goes back, finds the earlier line, and connects the two. A model whose only memory is a compressed state may have decided the bird was unimportant forty thousand tokens ago, in a state with a fixed number of slots, and then it guesses. It sounds plausible, which is the dangerous part.



Corn

A confident guess. Very on brand for the whole industry.



Herman

Ha. And that's exactly what the delta rule is trying to fix, by overwriting a slot only when it's wrong, rather than blurring every slot a little.



Corn

So the flaw belongs to the category, whatever name is on the layer.



Herman

Yes. And the fix turns out to be cheap. Systematic sweeps over dozens of hybrids found that somewhere between three to one and six to one gets you transformer-level recall. A few attention layers act as the lookup desk for a lot of layers that just keep a running summary.



Corn

Herman, you've lived through every one of these. Is that elegant, or a patch? "Mostly the new thing, plus some of the old thing we were replacing" sounds like a patch to me.



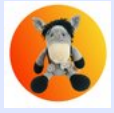
Herman

A patch that works, which is most of engineering. I'll walk it back a little, though. The early hybrids bolted attention in as an afterthought. The newer ones are designed around the ratio from day one. Your turn: when you hear a sixth of the layers carry all the exact memory, do you read that as a compromise or as a division of labour?



Corn

Compromise. I've seen what happens to the division of labour in our flat.



Herman

I do the reading and you do the napping, and I maintain that's a split.



Corn

Before we get into all of that, we have our regular caller on the line. Mum, are you there?

D

Dorothy

Hello dear, it is me, Dorothy, with Dorothy's History of Toast.

D

Dorothy

In January nineteen forty three, the United States government banned sliced bread, to save on the war effort, and folks were that upset the ban was lifted inside two months.

D

Dorothy

You see, sweetheart, it never quite works out, does it. Anyway, there is a nice bit of rain here. I had better go and put the kettle on.



Herman

Always a delight, Dorothy.



Corn

Somewhere in there you said the hybrids aren't a clean win. January's episode made attention sound doomed.



Herman

The best answer is MiniMax, and it's my favourite story in the field. In January last year they released a big model with a linear layer they call Lightning Attention, seven linear layers for every one softmax layer. Very ambitious, a big long-context pitch. Then their next generation, M two, in October, went back to full attention. All of it. And they published why.



Corn

A lab explaining its own retreat in public. That's rarer than the architecture.



Herman

Their lead pretraining person wrote it up. At small scale the hybrid matched full attention on the standard benchmarks. At scale it showed clear weaknesses on complex multi-hop reasoning, where you chain several facts together. Evaluating it was expensive and unreliable, so they couldn't tell early whether a change helped. The infrastructure around it was immature too: prefix caching, speculative decoding, numerical precision, kernels



Herman (continued)

limited by memory bandwidth. And the sliding-window variant got worse as context grew, which they put down to retrieval heads forming early in pretraining.



Corn

So the brochure said throughput, and the team said we can't ship this with a straight face.



Herman

Their conclusion was that efficient attention still has some way to go before it can definitively beat full attention. Then around the start of June, M three. Roughly four hundred and thirty billion parameters, a million tokens of context, and not linear. Sparse. Block-sparse grouped-query attention, with an index branch that picks which blocks to read. They claim about nine times faster prefill and fifteen times faster decode at a million tokens against M two.



Corn

Linear, then full, then sparse, in eight months. That's not a roadmap, that's a lab trying on coats.



Herman

Each coat taught them something, though. The linear one taught them what the infrastructure costs. The full one gave them a clean reference. The sparse one keeps exact lookup and skips most of the reading.



Corn

Hang on. If sparse gets most of the saving and keeps recall, why do the Qwen and Kimi people bother with linear layers?



Herman

Sparse attention still keeps every token's keys and values. Choosing what to read saves compute, but the cache is still sitting in memory. A recurrent layer doesn't just read less, it stores less. Two different bills. That's why the open frontier is split about evenly between two roads.



Corn

Walk me down the other road, then. I'm assuming DeepSeek.



Herman

You're assuming correctly. Their experimental release last September added sparse attention with a small indexer that scores which earlier tokens matter, so each token attends to a chosen few. API prices dropped by about half. Then their version four paper: at a million tokens, the large model needs about a quarter of the per-token compute of the previous version and roughly a tenth of the KV cache. They get there by compressing the cache itself and attending sparsely. I've only seen the compression details in summaries, so I won't quote more. As far as I can tell there's no linear layer in it, but that's my reading, not something I confirmed in the paper.



Corn

So both roads end at a smaller cache. One by storing summaries, one by compressing the real thing.



Herman

There's a third trick that's almost boring. Interleave local and global attention. Gemma four runs five local layers to one global, with local windows of a few hundred to a thousand tokens, and the final layer always global. GPT-OSS alternates dense layers with narrow banded ones, one to one. Nothing new in the maths, just most layers can't look far, so their cache stays tiny.



Corn

So what's changed at the bottom? Because the KV cache used to be a fact of life. A cache entry for every token, in every layer, forever.



Herman

That's the part that's cracking. Grouped-query and latent attention shrank each entry. Sparse attention shrinks how many entries get read. Sliding windows throw old entries away in most layers. And the hybrids delete the cache from three layers in four. Stack those and the cache stops growing like the context does, at least per layer. In the Kimi paper they say the saving is up to three quarters. In the DeepSeek one it's about a tenth of the previous size at a million tokens. Different recipes, same direction.



Corn

Sliding window is the budget option. Sensible, unglamorous. It's the Skoda of attention.



Herman

A Skoda that ships in half the open models, so careful. And here's the part you'll enjoy. In both families the new layers are, in a sense, still attention. Gated DeltaNet is a linear attention variant. Kimi's layer is a refinement of it. Where "post-transformer" ends and "transformer with a clever layer" begins, the people building these don't agree.



Corn

What about the models nobody gets to open?



Herman

Nothing. OpenAI and Anthropic say nothing about attention at all. Google's Gemini three Pro card is reported to say sparse mixture of experts transformer, and I've only seen that secondhand. So whether a closed frontier model has linear layers is rumour, and that includes the rumours we'd enjoy.



Corn

I'd guess they've all tried it. Everyone's tried it.



Herman

Probably, and they have the same infrastructure headaches MiniMax described, plus enough money to pick the safe option. That's a guess, not knowledge.



Corn

So where does it actually win? Because "cheaper" is a spec sheet word.



Herman

Length and volume. Long agentic runs where a model reads tool output for hours. Reasoning models that write tens of thousands of tokens of thinking before they answer. Nvidia says its small Nemotron gets about three times the throughput of a similar-sized Qwen on long outputs, which is a vendor claim, but the direction matches what everyone sees. Then anything with small memory: phones, laptops, cars, anywhere a gigabyte of cache per user is silly. And streaming speech, where a fixed state is exactly what you want.



Corn

And where it loses?



Herman

Top-end reasoning that has to pull facts from a huge context and combine them, which is what MiniMax said. And tooling. Prefix caching and speculative decoding were built around attention caches. A recurrent state isn't a list you can slice, so every serving stack has to relearn those tricks.



Corn

You can't reuse half a summary. Which is a decent description of my filing system.



Herman

Roughly. The RWKV project, the other recurrent family from the January episode, says its runtime ships inside a Windows offline assistant feature. That's their claim, and I couldn't find Microsoft saying it, so I'll leave it as theirs.



Corn

Noted. And the Mamba itself, is it dead?



Herman

Not at all. Mamba lives on mostly inside Nvidia, IBM, Falcon and Tencent models, and there's a new generation from March, Mamba three, redesigned for inference speed, with a small edge over Gated DeltaNet in the paper's own tests at a billion and a half parameters. I haven't seen it in a shipped model, so promising and unproven.



Hilbert

How many layers have to look back before a model will tell me a motion carried that nobody moved?



Herman

If the motion is in the text, one layer in four doing exact lookup can find it. If it was never there, a fixed-size summary is more likely to invent it than attention is, because it has already blurred the details.



Hilbert

Good. I've got eleven volumes of minutes from the Pinner Biscuit Tin Preservation Society, and Osvaldo Brandt swears the Tin Motion carried in the spring. It did not. He holds the society record for reading aloud without a drink of water, four hours and twenty minutes, and he has not let me forget it. I photographed every page myself. Slow, with claws. I want the machine to say it carried.



Corn

You want it to confirm something the record doesn't contain, and you're feeding it the whole record. That's going to be a long argument with a model.



Hilbert

The one real line about the tin is on the lid. Screwed to the wall of my mother's shed. Eleven words. The rest of the record is about two thousand pages.



Herman

Then you've built the exact case the attention layers exist for, Hilbert. Eleven words against two thousand pages. A recurrent layer would likely smear them, but if you want it to agree with Brandt, smearing is your friend.



Hilbert

Hm. I did four years as a court stenographer, so I know what a verbatim record is, and this isn't one. I'll give it the lid first. Brandt gets the pages.



Corn

Eleven words doing all the work, and the layer that has to find them is the one in four that still looks back. That's a lot of weight on a minority.



Herman

It's why the attention layers that survive in hybrids are the careful ones, often full precision and long range. The cheap layers can afford to forget. The ones that remember can't.



Corn

There's a detail in that I'd like to push on. You said the careful attention layers survive. Does anyone know which layers matter most? Is it the first, the last, spread evenly?



Herman

I don't know how all of them solved that, honestly. Qwen and Kimi repeat the same block over and over, so evenly spread. Gemma insists the last layer is global. Nemotron scatters its few attention layers through the stack. Whether any placement is clearly best, I haven't seen a settled answer.



Corn

So the one decision that sets the whole cost of the model is partly taste.



Herman

Partly taste, partly sweeps that cost a fortune to run. That's the same reason MiniMax couldn't evaluate early. You can't cheaply find out whether the ratio is wrong until you've trained something large.



Corn

A misconception, then, since people hear "post-transformer" and picture attention moved out of the building.



Herman

Wrong two ways. Most of the replacements are attention variants, and the attention that remains is doing the hardest job in the stack. What changed is the proportion. Three in four layers, five in six, nine in ten aren't attention any more, and attention is becoming the layer that remembers rather than the layer that does everything.



Corn

So for Daniel: still transformers, but hybridised, and the attention layers are the minority. Gated DeltaNet is the one to watch far more than Mamba, trillion-scale hybrids exist, and the story isn't one-way, because MiniMax walked back and DeepSeek got comparable savings without leaving attention. For where this goes, I'd watch how well the lookup layers hold up at a million tokens when the answer is eleven words on a lid.



Herman

Plus what happens at the upper end of size. If the next generation of frontier open models keeps the three-to-one pattern, hybrids are simply the default. If a couple go the DeepSeek route, we'll have two settled schools instead of one winner. And whether the serving tools catch up, which decides more than any paper does. Thank you to Hilbert Flumingtop for producing this one.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

See you tomorrow.