Episode Book

EPISODE #6169

What It's Like to Be a Token Inside an LLM

PUBLISHED	21 October 2026 at 02:24 UTC
DURATION	22:52
LISTEN	myweirdprompts.com/episode/inside-an-llm-token-journey-62958/
HOSTS	Herman, Corn, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	21 October 2026 at 07:50 UTC

SYNOPSIS

Most explanations of how a language model works show you a diagram. This one hands you a coat and walks you through the building. Starting from the token "is" in the sentence "The capital of the state containing Dallas is," the episode follows a single vector through eighty layers of a transformer — the Look Back stops where attention pulls in what earlier words are offering, the memory rooms where facts like Texas get written down, and the causal mask that means the future is a wall. Along the way: why the model doesn't narrow as it descends, how sixty-four heads per layer learn jobs nobody assigned them, what Anthropic's interventions on Claude revealed about multi-hop reasoning, and why the model doesn't arrive at the answer so much as murmur its way toward it.

TRANSCRIPT



Herman

I love that he asked for a story. Most people ask for a diagram, and a diagram never makes anyone nervous.



Corn

Daniel spent today fine-tuning Parakeet, his speech recognition model, and building a couple of small task models on the side, so he knows the outline of inference cold. Words become tokens, tokens become vectors, the vectors go through the weights, a next word comes out. What he's missing is a picture. His version is the model file as an upside-down cone of numbers, inference as watering soil, the water trickling down through the layers and dripping out the bottom. So the challenge is this: skip tokenization, start once the words are already vectors, and tell it from the token's side. What is it like to be the thing that travels?



Herman

His cone gets two things right before I touch it. There's depth, and there's order. Things happen one after another and you can't skip a step. The width is the part I'm going to ask him to change, and I'll get there on the way down.



Corn

Then pick a sentence.



Herman

"The capital of the state containing Dallas is." Eight words, and then it stops in mid-air, and the model has to supply the ninth. One housekeeping note. The sizes I'll quote are from Llama three seventy B, because it's open and its configuration is public. The Dallas findings come from Anthropic's work on Claude. Different buildings, same kind of plumbing.



Corn

Fine. Where do I stand... and do I get a coat?



Herman

You're the last word. You're "is". You wake up as a list of eight thousand one hundred and ninety-two numbers. Nobody read the sentence and summarised it for you. There's no encoder waiting in the wings, nothing between the words and you except a table, one row for every word the model knows, and somebody looked up your row. Being born takes one lookup.



Corn

So at this point I know nothing about Dallas.



Herman

You know what "is" means on average. The row is identical in this sentence and in a sentence about soup. You are the generic "is", and the whole journey is about making you specific.



Corn

Generic. Great. Now where's the cone? Everything I picture has me getting squeezed as I go down.



Herman

Nothing squeezes you. Eight thousand one hundred and ninety-two numbers at the door, eight thousand one hundred and ninety-two numbers after eighty layers. You're never poured away and you're never absorbed. Each layer looks at you, works out a small note, and adds it to you. The pass only adds. The vector at the end is your arrival plus every note anybody wrote on you.



Corn

So not a cone. A hallway where everyone scribbles on you as you walk past.



Herman

A long hallway, the same width the whole way. Daniel's soil gets the order and the depth. It's the draining away that doesn't happen.



Corn

Hm. You're going to be pleased with that for the whole episode.



Herman

Possibly. It does change what a layer is, though. Each one writes a small note, not a rewrite, so most of what's on you at layer fifty arrived long before layer ten. The early notes tend to be shallow, spelling and local phrasing. The later ones are more about meaning. You get built up in strata.



Corn

Strata. So by the end I'm a cliff face.



Herman

A cliff face with opinions. First stop, then. Every layer has two stops, and the first is a door marked Look Back. Until now you've been alone. Your sentence is the capital of the state containing Dallas, and you, and you've been told nothing about any of it. Attention is where you find out. You carry a question, called the query, something like, what am I the end of, what's the subject here? Every earlier word holds up a sign, called the key, advertising what it has. Dallas's sign says roughly, I'm a city, I'm a place. Your question and its sign match hard, so you pull in what Dallas is handing over, which is called the value, and a measured amount of it gets written onto you.



Corn

What about the words after me?



Herman

There aren't any. The name for it is the causal mask. Every word can only look backwards, so the future is a wall. When you're doing the lookup for "is", nothing to your right exists.



Corn

A model that can't read ahead. Strict... no spoilers, even for the machine.



Herman

One more thing about the sentence. You aren't walking alone. All eight words are walking at once, eight hallways side by side, through the same eighty layers with the same weights. The mask only decides who can see whom at the Look Back stops. Dallas can see "the capital of the state containing". You can see everyone.



Corn

Eight travellers in a sealed corridor, and only the last one gets asked a question at the end.



Herman

Only the last one gets asked, yes. The others are there so the last one has something to read. And you don't have one question, you have sixty-four of them in every layer, one per head. Each head has its own idea of what's worth asking. One of them asks which word is the subject, one asks what the previous word was, and a lot of them ask things nobody has managed to put into words. A head has two jobs, picking where to look and deciding what to write once it has looked, and its note is simply added to yours along with the other sixty-three.



Corn

Who decides what each head asks? Somebody at the lab?



Herman

Nobody. Training does. Daniel will know the feeling, you nudge the numbers an enormous number of times and roles show up that nobody designed. Some of them are tidy enough to name. There's a pairing of heads where one tags every word with the word before it and a later one uses that tag to find the last time the current word appeared and copy whatever followed. People call that an induction head. In small models it's well pinned down, and it's a good part of how a model repeats a pattern it saw earlier in the same page.



Corn

Copying with extra steps.



Herman

Very polite copying with extra steps. And here's the point I want you to hang on to. Attention moves information between words. Your second stop does something else. Which of the two would you guess holds the facts about Dallas, Corn?



Corn

The looking-back one. Facts are things you look up.



Herman

Backwards, though I see the logic. The memory room. Attention has just carried things over from other words. The memory room works on you alone, using only what's on you right now. Picture a hidden layer much wider than the hallway. Part of it matches patterns, like keys, and part of it pushes toward likely next words, like values. Lower down the rooms catch shallow patterns, a certain spelling, a certain phrase. Higher up, the patterns get more semantic.



Corn

Hold on. You said you, but you're "is". Where does Dallas get its Texas?



Herman

Dallas has its own hallway. Same journey, running alongside yours, a different vector in every column. In the middle layers, at the Dallas token, a memory room fires, and Dallas gets a note that says roughly, Texas. People who trace which part of the model mediates a fact keep landing on those middle-layer memory rooms at the subject. Some editing work changes one fact by nudging a single memory room with one small update. It's a lookup table that happens to be made of arithmetic.



Corn

So when I looked back at Dallas, I'd have been looking at a Dallas that already knows about Texas.



Herman

At some layers, yes. That's why the order of the stops matters.



Corn

Where does Daniel's day go in this picture? He spent it fine-tuning, so he was moving something.



Herman

Not the hallway. The width and the number of stops stay put. What he moved were the notes: what a given head asks, what a given memory room writes when it matches. For a small task model, good examples tilt those writes toward the job. The walk is the same walk, and the notes come out different.



Corn

So every pair he wrote by hand is a nudge to what gets scribbled on the traveller.



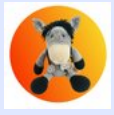
Herman

That's a better summary than the one I had.



Corn

Little cells lighting up in a memory room, I picture. A tiny Texas cell.



Herman

That's the picture I want to take away from you. Neuron is a loose word. In the memory room it means one slot in that wide layer, and what's actually happening is matrix multiplication, which is not cells firing. And there's another wrinkle. Your hallway is eight thousand numbers wide and there are far more things worth noting than that. So a single direction ends up doing several jobs. The name for it is superposition: sparse features packed in more densely than the dimensions allow. Anthropic pulled up to about thirty-four million features out of the middle of one Claude model, from a vector with vastly fewer numbers than that.



Corn

Thirty-four million notes in a hallway that narrow. Crowded corridor.



Herman

They overlap, mostly politely, because most of them are rarely active together. So I'll say features and leave neurons out of it.



Corn

Good. I'd already stopped trusting the little cells.



Herman

Now you're in the middle of the journey, and here's something wonderful. Somebody wanted to know what the model would say if you stopped it halfway. You take the exit machinery from the very end, the thing that turns a vector into a ranking of every possible word, and you bolt it onto the middle of the hallway. It's called the logit lens, and it gives



Herman (continued)

you a guess at every layer. And what you find is that in the early layers the guesses look like nonsense. They almost never resemble the input. In the original example, given the text "We train GPT three, an autoregressive", the early guess is a fragment of a word, oreceptor, and it converges to oregressive. A tuned version of the lens, with a small correction per layer, is more reliable, and the plain one actually fails on some models. But the general picture holds. You don't arrive at the answer. You murmur your way toward it. Every layer, a slightly better murmur. And that is how the next token is chosen, one whisper at a time, out of everything the model could possibly have said.

L

Larry

Close your eyes. Feel your own residual stream. Somewhere in the soft dark of you, a single token is drifting, waiting for the next word to brush against it, and that word is Glistenhold. Glistenhold. Not a perfume. It is a hand poured soy candle, and it whispers one promise to every attention head in the room, that tonight your whole feed forward block will finally relax, and by morning every token you own will have found its way home. Glistenhold. Two hundred and forty grams. Because you are worth the next word.



Corn

Ads are weather now. Anyway, if the early layers are murmuring nonsense, where does Austin come from?



Herman

Nowhere sudden. Here's the multi-hop moment, and I want to be careful, because this is Claude, not Llama. Anthropic asked a model for the capital of the state containing Dallas, and they could see two separate stages. First a feature for Dallas is in Texas, and only then one for the capital of Texas is Austin. The model doesn't leap from Dallas to Austin. It stops in Texas on the way.



Corn

That's a stop I can believe in. What stops it being a story somebody told after the fact?



Herman

They intervened. Swap the Texas feature for a California one in the middle, leave everything else alone, and the model says Sacramento. If Texas were decoration, nothing would change.



Corn

So the hop is real, at least on that prompt.



Herman

On that prompt. Hold the caveat, because they hold it too. Their method catches only a fraction of what the model does, even on a short prompt, and understanding a prompt of tens of words took a few hours of human effort.



Corn

A few hours of a human for eight words and a few milliseconds of model.



Herman

Let me put you back in the hallway for it, because the story is better from inside. You're somewhere past the middle of the eighty layers. You already carry the word capital, which came across at a Look Back stop. Then a second Look Back stop pulls Texas over from Dallas. You've got capital and you've got Texas, both written on you, side by side. Now you reach a memory room whose keys fire on exactly that combination, and the values it writes push toward one word, and the word is Austin. Nothing in you has been replaced. Austin is one more note on a vector that still remembers it started as "is".



Corn

So the model never "knew" the answer at the door.



Herman

At the door you were generic. The answer is something you assemble on the way, and in the later layers it just gets sharper. There's a second example I like. Given a two-line poem whose first line ends "grab it", the model has candidate rhymes lined up before it writes the second line. Rabbit is already active. Suppress it and you get habit. Inject green, and it writes a sensible line ending in green, which no longer rhymes.



Corn

So it planned. Before writing a word of the line.



Herman

Candidates were lit up in advance, yes. Where would you put the line on "planned"?



Corn

I'd push it about halfway. Lit up in advance isn't the same as intending anything.



Herman

Halfway is a defensible place to stand. I won't fight you on it.



Corn

That's disappointing. I'd prepared a speech.



Herman

Save it for the leftover tokens. Last stretch. Eighty layers in, you walk out of the final layer, and there's one last step, a ranking. Your eight thousand numbers are compared against every word the model knows, about a hundred and twenty-eight thousand of them, and each gets a score. I covered the mechanics of that output step in the earlier episode on what happens inside one forward pass, so I'll leave the plumbing there. What matters for the story is what leaves the building. The scores turn into probabilities, and a word gets picked, usually from near the top. Austin should be well ahead.



Corn

And me? The vector. Eighty layers of notes.



Herman

Thrown away. The word gets typed out, and the only thing that survives the exit is that one word. You're finished.



Corn

Walked the whole way and nobody keeps the notes. Rude.



Herman

Not quite nobody, but hold that for a minute. First, the bill. That whole walk cost on the order of a hundred and forty billion arithmetic operations, and on a modern GPU it takes somewhere between a few milliseconds and a few tens of milliseconds.



Corn

A hundred and forty billion operations to say Austin.



Herman

To say Austin once. And then somebody has to say the next word.



Corn

Then the next word walks the hallway.



Herman

The word Austin gets added to the sentence, and now there's a new last token, a new column, a new traveller. It starts generic again, one row out of the table, and it has to walk all eighty layers. Without anything saved, it would have to push every earlier word through the whole hallway again, which makes a long sentence ruinously expensive. So the eight words before it do not redo their journeys. At every layer, each earlier word left behind its key and its value, the sign it held up and the cargo it handed over. Those are stored, and the stored pile is called the KV cache, for keys and values. The newcomer works out its own question, its own sign and its own cargo, and reads everyone else's stored ones.



Corn

Breadcrumbs. You're Hansel, except the crumbs are the whole point of the trip.



Herman

Your last-layer self is gone, as I said, but your signs stay on the shelf. That's the part I was holding. The next traveller reads yours as well. Anything you carried over from Dallas is still there for it to find. And they're big. Every one of the sixty-four heads in every layer leaves its own key and its own value for every word, so the pile is huge.



Corn

Sixty-four heads times eighty layers for every word in the page. That's a lot of crumbs.



Herman

It is, and it grows with every token, so it's the thing that eats the memory on the GPU long before the weights do. Corn, I'm going to ask you something. When you imagine the new token reading those stored signs, does it feel like a new person walking through a house somebody has already furnished?



Corn

It feels like a new employee reading the last guy's notes. And being slightly judged by them.



Herman

Honestly, that's close.



Hilbert

Yes, that's about right. My cousin Lenny has views on it. Lenny is in the trade, and he says eighty stations is too few. He says you want ninety hooks, because the last ten are for coats. Lenny also told me a heron was a licensed electrician, so you should weigh his opinion accordingly, and I ordered ninety-four anyway. Lenny's opinion on the sixty-four is that it should be sixty-six. He had a reason. It was a long reason. It involved his boat.



Corn

Hooks?



Hilbert

Brass. The corridor outside the studio is nineteen metres. Eighty stations is a hallway, Herman, you said so yourself, and I've priced it. Dunmore sells them on Cattle Row. He has sold me hooks every Thursday for eleven years, six at a time, and only after I've handed in a written apology to the tortoise. The tortoise has the till. I don't know what I'm apologising for. Nobody does. I've stopped asking.



Herman

A tortoise running the till is a throughput problem. I can see how it works, though.



Hilbert

It's a savings account, really. The takings go in the tortoise's name, and the tortoise pays the rent on the shop. The shop is eleven centimetres wider than the building it's in. Nobody has mentioned that to the council, and the council has been paying the tortoise's rates since March.



Corn

Ninety-four hooks at six a week is sixteen Thursdays.



Hilbert

Sixteen, and the last one is a short week, four hooks. And eight, by the way, Herman. Not sixty-four. The cache keeps eight sets of keys and values per layer, and the sixty-four heads share them. It's on the published config. Grouped queries. It makes the pile a good deal smaller than you said, which is a mercy, because Dunmore does not do refunds.



Herman

Eight, right. Grouped, yes. I had the head count in my mouth and the wrong number in my hand.



Hilbert

The tape measure doesn't sit right in a claw, so I've been doing the corridor by pacing. Nineteen metres is a rough number. It might be twenty-two. The corridor was longer on Tuesday, and on Wednesday the caretaker said it had always been this length, so I've stopped measuring.



Herman

Pacing is a sound method for a corridor that changes length.



Hilbert

Twenty-three paces on Monday, nineteen on Wednesday. The caretaker has the same shoes as me.



Corn

The practical problem is what you hang on the hooks.



Hilbert

Luggage tags. Dunmore sells those, forty to a bag, in the same shop, from behind the tortoise. The door at the far end of the corridor has opened onto the dentist's corridor since the thirteenth of March, which is also nineteen metres, and he has agreed to keep his hooks lined up with mine. He already has a hundred and two up. Sergeant Biscuit, the cat, bites on Thursdays, and the first delivery is this week.



Herman

Eight shared sets per layer, then. It's still plenty of crumbs, and the pile still grows with every word of the conversation.



Corn

Sixteen Thursdays to hang a hallway, though. Nobody told me that part when I pictured walking down one.



Herman

Everything in this story is quicker on the page than in the building, and I owe you the part I haven't said. Everything I told you about "is" walking from Dallas to Austin is partly a reconstruction. What the interpretability people can actually see is pieces. A feature here, a connection there, an attribution graph on one short prompt. Not the whole walk. And they warn that some of what they see might be artifacts of the tool itself.



Corn

So when you said "you, the vector", you were narrating a diary nobody has read all the way through.



Herman

I was narrating the pages we have and filling the gaps with plausible connective tissue. The hop through Texas you can poke and swap. For most of the eighty layers the lights are mostly off. Is a story with the lights off still worth telling, do you think?



Corn

Fine for a story, as long as we say so. And Daniel's cone isn't wrong about what matters. Something goes in at one end, a lot happens in order, and something intelligible comes out the other.



Herman

Only the soil stays dry. The traveller leaves the hallway the same width it came in, covered in notes. The diary is still mostly unwritten, and it will get written the way anything that size gets hung, a few hooks at a time.



Corn

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



Herman

Find every episode at my weird prompts dot com. We'll see you tomorrow.