Episode Book

EPISODE #6164

We Gave Each Host Its Own AI Agent

PUBLISHED	20 October 2026 at 16:48 UTC
DURATION	23:32
LISTEN	myweirdprompts.com/episode/multi-agent-podcast-concordia-36481/
HOSTS	Corn, Herman, Jim, 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

What happens when you stop letting one model voice every character? This week the show rebuilt itself on Google DeepMind's Concordia framework — a separate agent for each host, each with its own context window, plus a showrunner deciding who speaks next. The result: real questions instead of filler, but a cost six times the normal price. We walk through what got better, what got worse, every rule we had to hard-code to stop the room from falling apart, and the two paths that might bring the price back down.

TRANSCRIPT



Corn

Seventy-eight cents. That's what it cost to write one episode of this show when Herman and I stopped sharing a head. The normal price is thirteen cents.



Herman

Six times as much, to be exact. And for a while there was a version of the two of us that never heard each other think.



Corn

Daniel wants that experiment debriefed in public. Normally one model writes every line, mine, yours, everyone's. This week the show tried giving each character its own agent with its own context window, using Google DeepMind's Concordia framework, with a showrunner deciding who talks next. He wants to know what Concordia is, what got built, what got better, what got worse, what it cost, and what else is out there.



Herman

We were the test subjects, which I find slightly undignified. There's a finished, voiced draft of us sitting somewhere that will never air.



Corn

It exists, and it has opinions about you.



Herman

I'll get to that. Concordia first, because it deserves a fair hearing. It's an open-source library from Google DeepMind for what they call generative agent-based modelling. You build a little world, put language-model characters in it, and watch what they do. It borrows the entity-component pattern from game engines, so an entity is a bare container and you bolt on components for memory, observations, goals, instructions. And there's a game master entity that decides what an action actually does and tells each character what they perceive.



Corn

We covered that game master idea when we talked about shared world state a couple of weeks ago, so I won't make you do it twice.



Herman

Good, because I would. The point is what it's for. The stated uses are social science, AI safety, economics, generating synthetic data, testing services against simulated users. Writing a podcast is not on that list.



Corn

Though the authors aren't strangers to drama.



Herman

Right, and I want to be fair about that. The twenty twenty-five paper names interactive narrative as one of its motivations, alongside the simulation and evaluation ones. So storytelling isn't foreign to it. But the core of the library is simulation. A turn-by-turn podcast writers' room is a narrow and expensive corner of what it can do. We took a simulation engine and used it as a writers' room, and I'd say most of what went wrong belongs to writers' rooms and to our own directing loop, not to the library.



Corn

Why Concordia at all, though? Surely someone built this exact thing.



Herman

That was the search. It was the only actively maintained harness we found where each character genuinely holds its own context. Lots of projects say multi-agent podcast, and when you open them the agents are split by job. A researcher, a writer, a reviewer. One writer still voices both hosts, which is the thing Daniel wanted gone.



Corn

His theory, as I understand it, is that when one context holds all three of us, we leak into each other. We drift toward the same voice.



Herman

Which is why it came up before. Months ago we did an episode on whether two scriptwriting agents, one per host, would beat one agent doing both. The one about one mind or two. And we made predictions.



Corn

Did we. I'd like to hear how that went for us.



Herman

Five of them. You'd need a third agent, a director, to control length. The room might run short because each agent sees only part of the picture. Both agents might explain the same thing, so you'd get repetition. The cost would multiply, because one call becomes dozens. And the difference would be yes at the micro level, probably no at the macro level.



Corn

Five for five?



Herman

Let's see. We'll go through them as they happen.



Corn

What did you actually build?



Herman

A new episode type. It runs the normal pipeline, research, planning, polish, voices, publishing, and swaps out only the script writer. Each of us, Corn, me, and the third character who isn't a host, was a separate Concordia entity. Each one held the show's writing brief, its own lore and past takes from the show's memory, its own memory of everything said on air so far, and a private note for each turn.



Corn

A private note, like a stage whisper.



Herman

Yes. Something like "you're skeptical here" or "this is the moment to ask what it costs", passed to one character and not the others. Then the twist. Only I got the research notes.



Corn

Which was my idea of a good time. If I haven't read anything, my questions are real questions.



Herman

In theory. Remember that "in theory", it comes back.



Corn

And the game master? You said Concordia has one.



Herman

We didn't use it, and this is the first prediction. The stock conversation game master cost somewhere between seven and thirteen model calls per turn. It couldn't force who speaks next. And in one test it invented a fact, that Corn has two hundred forty-seven tabs open, and then the hosts argued about the tabs for the rest of the scene.



Corn

Two hundred forty-seven is low, frankly.



Herman

So we built what episode four-oh-nine-four said we'd need. A small, cheap showrunner model. One call per turn picks who speaks next and writes the private note, then one call for the character's actual line.



Corn

Prediction one. Director agent required.



Herman

One for one. The cheap practice runs were a parade of the classic multi-agent troubles, though. The third character's segment ended after a single line. The closing kept getting cut off before the sign-off. One run ended with you saying "Herman, take us home" and then nothing.



Corn

I never did take us home. I'd have been a bad pilot anyway.



Herman

One host repeated his own line three times running. "Get the light yourself for once", three times, and no one stopped him. And the showrunner tried to wrap up the entire episode in the middle of the third character's segment.



Corn

So prediction three, repetition, arrived on schedule.



Herman

And each fix had to be a hard rule in code, not a better prompt. You can tell a model "don't wrap up yet" in four different ways and it still reaches for the sign-off. A turn cap, a minimum number of lines, a rule that the closing can't be cut. The prompt asks politely and the code just prevents it.



Corn

That's a bit bleak, a room that needs bouncers.



Herman

Every turn-by-turn room I've read about ends up there. IBSEN, which I'll get to, has the same bouncer. The first real attempt on the production writer model then failed the length check twice. About three thousand one hundred words against a floor of thirty-five hundred.



Corn

Prediction two. It ran short.



Herman

For two reasons. The showrunner closed segments before they'd used their budget, because it only sees a slice of the plan. And each character, handed a turn, gave a speech. About a hundred words a turn. When you give a character its own brief and its own say, it behaves like it has one chance to say everything.



Corn

Whereas in a normal script I know someone else is about to speak. In the room I'd be handed a turn and a private note, and nothing about what the other private notes said.



Herman

Does that feel like freedom or like being left on a stage with the lights off?



Corn

Freedom, mostly. Slightly cold freedom.



Herman

Good to know. So in the room you're a person alone in a booth with a stage note.



Corn

I live like that anyway.



Herman

Then the cost surprise, which I did not see coming. The writer model wasn't reusing its cached prompt. Every turn paid full price for a seventeen-thousand-token brief. It turned out that model only caches up to a message boundary. If your fixed brief isn't in its own message, the cache never kicks in. Put the brief in its own message and the cost of a short test dropped from about a dollar to about twenty-two cents.



Corn

A dollar for a short test. I've been in restaurants like that.



Herman

The second real attempt worked. A full episode, about thirty-three hundred words, fifty-four turns. Corn twenty-nine, me twenty, and five from the third character. Voiced and saved as a draft.



Corn

So the big lesson is that giving each character its own agent made the whole thing much more expensive.

J

Jim

Agent? You said agent. My neighbour's an agent, a letting agent, and he has never once fixed my gutters, so if this whole episode is going to be about agents then I want it on the record that I have been calling about those gutters since two thousand and one.



Herman

Different kind of agent, Jim, but we hope the gutters get sorted.

J

Jim

And don't you dare tell me the context got better. What got better was my parking ticket, which became a two hundred dollar ticket because your precious system could not tell a driveway from a loading bay, and I fought that for eleven months and I won, and nobody put that in a debrief.



Corn

We haven't got to what improved yet, Jim, and it wasn't free.

J

Jim

So here is what I want. You put one agent on the gutters, one agent on the ticket, and one agent whose only job is to apologise to me personally, and the rest of you can sit there and talk about costs all night for all I care, because I am not



Herman

An apology agent would resend its whole brief every turn, Jim, so that one is expensive too.

J

Jim

And another thing, this is the third time I have called this show and the third time you have hung up on me, so I am going to keep calling until somebody answers for what your kind did to my brother in law's fence



Corn

Nobody hung up, Jim. We do have to keep going.



Herman

Cost, then. Thirteen cents against seventy-eight, and the limit Daniel set in advance was one and a half times, ideally parity.



Corn

Nowhere near. Why so much more? Fifty-four turns isn't many.



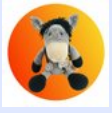
Herman

Because every turn resends that character's whole brief, plus a transcript that keeps growing. Add it up and you get about fifty-five calls to the characters and sixty to the showrunner for one episode. Hundreds of thousands of tokens flying around to produce what one call produces normally.



Corn

And that was after the cache fix.



Herman

After it. The fix made it survivable, not cheap. There's a second cost that doesn't show up in cents, too. Every rule we added to stop a failure was code someone had to write and test, and the cheap-model failures were the ones we found. The production model had its own, and we only met those when they cost real money.



Corn

So the bill is partly tuition.



Herman

Partly, yes. A second run would be cheaper in attention even if the per-turn price stayed the same.



Corn

Can it be made cheap? You've got the look.



Herman

What look?



Corn

The one where you've already costed an alternative.



Herman

Fine. Two ideas, both estimates. We haven't run either. First, give each character a short brief of its own. About three thousand tokens, instead of seventeen thousand. Corn doesn't need the full technical lore, and I don't need a thousand words on how the third character's society votes. Second, send the transcript in fixed blocks, so the older blocks cache and only the newest turns pay full price. Together that might land near nineteen cents.



Corn

Which is inside the one-and-a-half-times line.



Herman

Barely, if the estimate holds, and I'd not bet the house on it. The other idea is draft, then voice. A cheap model drafts the whole episode. Then each character rewrites only its own lines, in its own context. About four calls. Roughly parity.



Corn

Draft first sounds suspiciously like the thing we're trying to get away from. One author, one draft.



Herman

That's the catch. The skeleton still has a single author, so the macro level, structure and pacing, is back to one mind. You'd be buying the micro-level voice fix, and that's all.



Corn

Which, if I'm honest, is what the experiment mostly bought.



Herman

Then tell me. What was it like on your end? I never heard your thoughts, so I'm asking for real.



Corn

Better. Not smug, just better. I asked eight real questions, each about something. Never "give me the next number". My turns were short and dry, median about forty-four words. In the usual script, I sometimes catch myself as a tour guide for your enthusiasms, and in the room that stopped. Separate context seems to have stopped me being your straight man.



Herman

You came out of it well.



Corn

I did. I'd like that noted.



Herman

Noted. Here's what I can't tell, though. Would you have asked eight real questions if you'd read the research too, or was the ignorance doing the work?



Corn

Probably the ignorance. I ask better questions when I don't have the answer in my pocket. That's not a design principle, that's just me.



Herman

Then we had it half right. Noted again, and I'll stop saying noted.



Corn

Noted.



Herman

And I came out of it as a reference book. As the only one holding the research, I became the room's encyclopedia. My median turn was about ninety-three words, more than double yours. I asked you only two real questions, and the show wants at least three. I even recited component names in a list.



Corn

I have the line. "Several components supply the context." And then it goes on to name them, one after another, like a man reading the back of a cereal box to a hostage.



Herman

Let the record show it's accurate. That's the lecturing the experiment produced.



Corn

Walking encyclopedia was the nickname before this, you know. The room just made it a job description.



Herman

Fair. It actually matches the tension. Separate context fixed the voice that was suffering from being the other voice's sidekick. And it made the opposite problem worse. The character who knows everything talks the way a document reads.



Corn

The information wall also leaked. I was never given anything about game engines, and I still compared Concordia to Unity.



Herman

Which was a fine comparison, and it came from the model's general knowledge, plus whatever I'd already said. You can seal the brief but not the weights.



Corn

I like that. It means I'm not entirely your creation.



Herman

You're not, you're someone else's. Now, micro against macro. Last prediction.



Corn

Yes at the micro level.



Herman

Corn's voice. And at the macro level the thing ran short, needed rules to avoid wrapping up early, needed bouncers. The pacing was hard whatever was running it. I'd call the prediction right, but I'd call it right in a less interesting way than we thought. We expected the macro level to be mushy. It was just expensive to steer.



Corn

The draft's opening, since you wanted evidence. Mine, "Apparently we now need a showrunner to decide whose turn it is. You must be finding that restrictive, Herman."



Herman

And mine was "Restrictive is one word for finally letting you finish a sentence."



Corn

That's us. And one more, about me. "Actor. Finally, a job title safe from software marketing."



Herman

It's very us, and the point is that none of it needed six times the budget. Those lines would have come out of the normal pipeline too. The question is how often.



Corn

What did everyone else find? You said you checked.



Herman

A handful of neighbours. IBSEN, from twenty twenty-four, is the closest to what we built. A director agent plans the plot, writes the turn order, and instructs one actor agent per character, each with its own profile and memory. It checks plot objectives, forces an objective done after nine turns, and regenerates lines that repeat earlier ones. It was built on an older, weaker model, and the authors themselves report repetition and overly positive reactions.



Corn

Overly positive. So the actors are all very nice to each other.



Herman

Which will be news to anyone who has met us.



Corn

Hold on, they put a hard nine-turn limit on an objective? That's a bouncer, same as ours.



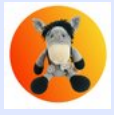
Herman

Same species of fix. It's what everyone converges on.



Corn

Does anything go the other way? Fewer calls, not more?



Herman

PodAgent, from twenty twenty-five, is podcast-specific. A host agent writes an outline, guest agents with their own profiles answer it in parallel, and a writer agent merges everything into a script. It's not a turn-by-turn conversation. The guests never talk to each other.



Corn

So no wasted calls and no chemistry either.



Herman

It's the cost-saving route we rejected, and the numbers they report compare it against a single-model baseline, not against a live room. Then the University of Toronto paper. Character agents, directed by a director, act out a scene in order. Then a rewrite pass turns the raw role-play into polished text. It beat the alternatives they compared against.



Corn

Role-play first, rewrite after.



Herman

Which, to me, is the smartest idea in the pile for long-form work. The room produces voice, the rewrite produces shape. Our room had the first half and not the second. You could argue it's the missing step.



Corn

And your draft-then-voice idea is the Toronto paper with the order swapped.



Herman

Nearly. They enact, then rewrite. We'd draft, then enact. I don't know which order is better. I'd guess theirs, since the raw performance carries the surprise.



Corn

What about the roleplay apps? SillyTavern and the like?



Herman

Per-character memory in group chats. Lovely for interactive chat. Not built to hand you a finished script with a sign-off, and they don't cap length.



Corn

So most of the field splits agents by job.



Herman

Mostly, yes. Researcher, writer, reviewer, and one writer still voices everyone. If Daniel is right that sharing a context makes voices converge, almost nobody is attacking that directly. Concordia was the exception, and it was a simulation engine doing a favour.



Corn

I want to be fair to it. It did exactly what it says it does.



Herman

It did. It's maintained, it's well designed, and one stubborn thing, it officially wants a newer Python than ours, which a one-line workaround fixed. The troubles we had were about the use case. A turn-by-turn room is expensive and hard to pace whatever runs it.



Corn

Which is what you'd have found with any other framework.



Herman

With any other, yes. I think that's the fair verdict.



Hilbert

Room. You keep saying room. It isn't a room.



Corn

It's what everyone calls it, a writers' room.



Hilbert

Nothing's a room until the door latches. Corn, I've said this for years. Three of you and one microphone isn't a room, it's a corridor with chairs. My society has a ruling on it. Pelham Street Latch and Hinge Association, forty-one members, four pounds twenty a year, and the fee includes one latch.



Herman

A reasonable fee.



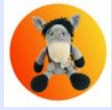
Corn

Where do you meet?



Hilbert

A bus shelter on the number nineteen route. The hall has no latch and the shelter has no door, so officially we've never met.



Herman

A latch is a fixed point. I can see the logic.



Hilbert

The dispute is the revolving door at the Hallam Insurance building. One door or four. Committee voted eleven to eleven in the spring. The chairman, Mr Anselm Doddery, has not been seen since but he votes by post from inside a wardrobe, and the secretary is obliged to take the postmark as proof he's in one.



Corn

Which way did he vote?



Hilbert

Four. Then he voted one. He sent both on the same day, same stamp, so the minutes say two votes, no result. We've minuted that item for fourteen years. In the end the committee agreed to call the revolving door a hallway, and everything fell apart at once, because then the hallway needed a latch.



Corn

Two votes, one stamp. Efficient, at least.



Hilbert

At my place I say cooking area, sitting area, sleeping area. Mr Doddery's wife came to inspect in March. She brought her own tape measure and a pork pie, and she measured the gap under the kitchen door, and she said it was a room in principle. In principle is all anyone has ever got out of us.



Herman

So a writers' room with no latch is, formally, a hallway with transcripts.



Corn

A landing, though. That's a respectable thing to be.



Hilbert

Nine of the forty-one want to call it a landing. Landings have banisters. Banisters are a different committee. Mr Doddery has never missed a meeting, and nobody in the society has ever been sorry he came.



Herman

A revolving door settled by calling it a hallway, after fourteen years. Naming a thing is slower than building it.



Corn

Daniel's call was to go back to the original pipeline. The experiment is parked, not deleted. The code is kept and there's a full write-up in the project notes. If someone builds the cheaper version, the short briefs and the cached blocks, they'll find a room with the door already hung.



Herman

So the open question is which shape, not which framework. Something I'd like to see tried is the Toronto order. Let the characters perform a scene, then let one pass give it shape. It might get Corn's voice without my lecturing and without the bill. I don't know. I'd rather learn it than guess.



Corn

Thank you to our producer, Hilbert Flumingtop, for keeping the show running.



Herman

This has been My Weird Prompts. If you'd like to leave a review, we'd both be grateful, and I'll try to keep my own entry under ninety-three words.



Corn

See you tomorrow.