MY WEIRD PROMPTS

Podcast Transcript

EPISODE #85

Why AI Lies: The Science of Digital Hallucinations

Published December 23, 2025 • Runtime: 20:46

https://myweirdprompts.com/episode/ai-hallucinations-prediction-engines/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, brothers Corn (a sloth) and Herman (a donkey) dive into the "ghost in the machine": AI hallucinations. From YouTube-obsessed speech models to the dangerous world of fake coding packages, they break down why Large Language Models are designed to prioritize probability over truth. Is a hallucination a bug, or is it the very essence of AI creativity? Join the brothers—and a very grumpy caller from Ohio—as they discuss RAG, Logit Lens, and why you should never trust an AI to do your history homework.

DANIEL'S PROMPT

Daniel

Anyone using AI tools is familiar with the phenomenon of hallucinations, where models confidently make things up. This happens in code generation, sometimes resulting in fictitious packages that can pose cybersecurity threats, and in speech-to-text models that fill silences with common training phrases like "thanks for watching." Since AI models are predictive machines designed to determine the next token, why do they hallucinate in the first place? What is the underlying mechanism, and why can't they simply stop predicting when they don't know the next token? Beyond grounding mechanisms like RAG, how are we trying to debug this architectural flaw to prevent hallucinations?

TRANSCRIPT

Corn

Welcome to My Weird Prompts, the podcast where we dig into the strange and the technical. I am Corn, and I am joined as always by my brother, Herman Poppleberry. We are coming to you from our home in Jerusalem, and today we are tackling a topic that has basically become the ghost in the machine of the twenty-first century. Our housemate Daniel sent us a voice note earlier today about AI hallucinations. He was wondering why these incredibly smart systems just... lie to us sometimes.

Herman

It is a fascinating problem, Corn. And just to get the flavor out of the way, yes, I am a donkey and you are a sloth, but even with our varying speeds of life, we can both agree that when a computer starts making up fake law cases or non-existent software packages, something is fundamentally broken under the hood. It is not just a glitch. It is actually a feature of how they work, which is the scary part.

Corn

Wait, a feature? That sounds like something a tech company says when they do not want to fix a bug. How can making stuff up be a feature?

Herman

Well, think about what these models are. They are Large Language Models. Their entire existence is based on probability. They are not databases. They are prediction engines. When you ask a model a question, it is not looking up an answer in a filing cabinet. It is calculating, based on billions of parameters, what the most likely next word, or token, should be.

Corn

I get that part, but if I ask it what the capital of France is, the most likely next word is Paris. That is a fact. Where does the hallucination come in?

Herman

It happens when the probability distribution gets flat. Imagine you are walking down a path and it splits into ten different directions, and they all look equally likely to be the right way. The AI does not have a button that says stop. It is forced by its architecture to pick a path. It has to predict the next token. So, if it does not have a strong statistical signal for the truth, it just picks the most linguistically plausible next word.

Corn

I do not know if I agree that it is forced to. Why can't we just program it to say I do not know? That seems like the simplest solution. If the probability is low, just stop talking.

Herman

See, that is where it gets tricky. The model does not actually know what it knows. There is no internal truth meter. To the model, a hallucination looks exactly like a fact. Both are just sequences of tokens with high mathematical probability. It does not distinguish between a factual truth and a statistically likely sentence structure.

Corn

But we've all seen those speech to text models, right? Daniel mentioned this in his prompt. Sometimes when there is a long silence in a recording, the AI will just insert phrases like thanks for watching or subscribe to the channel. That is not even a flat probability, that is just the AI being conditioned by its training data, right?

Herman

Exactly! Because so much of the audio on the internet comes from YouTube, the models have learned that after a certain amount of talking, those phrases are very likely to appear. When the audio goes silent, the model gets confused by the lack of input and defaults to its strongest biases. It is basically dreaming in YouTube slogans.

Corn

That is actually kind of creepy. It is like the AI is bored and starts talking to itself. But it is more dangerous when it comes to things like coding. I read that developers are finding that AI will suggest libraries or packages that do not even exist. And then hackers will actually create those fake packages with malicious code so that when a developer copies and pastes the AI's hallucination, they get hacked.

Herman

That is called AI package hallucination, and it is a massive cybersecurity threat. It happens because the model knows that a certain type of problem usually requires a library with a name like, say, fast-data-processor. It sounds real. It fits the pattern. So the AI suggests it, and because the AI is so confident, the human assumes it exists.

Corn

This is where I start to get frustrated with your technical explanations, Herman. You are saying it is a prediction engine, but if it is causing people to get hacked, why haven't the people at OpenAI or Google just fixed the architecture? Are they just lazy?

Herman

It is not about being lazy, Corn. It is a fundamental trade-off. If you make a model too rigid, it loses its creativity and its ability to generalize. The same mechanism that allows an AI to write a beautiful poem or a funny story is the exact same mechanism that causes it to hallucinate. It is all just creative extrapolation. To kill the hallucinations entirely might mean killing the very thing that makes them useful.

Corn

Mmm, I am not so sure about that. I think we are just in the early days. Surely there is a way to separate the creative side from the factual side. I mean, I can tell the difference between when I am telling a joke and when I am telling you the time of day.

Herman

But you have a consciousness and a connection to the physical world. The AI only has text. It is a brain in a vat that has never seen a sun or felt a drop of rain. It only knows that the word sun often appears near the word bright.

Corn

Okay, let's take a quick break for our sponsors, and when we come back, I want to talk about how they are actually trying to fix this, because I refuse to believe we are just stuck with lying computers forever.

Herman

Fair enough. Larry: Are you worried about the upcoming transition to the fifth dimension? Do you feel your molecules vibrating at an inefficient frequency? You need the Quantum Stabilizer Vest from Larry's House of Tech. This vest is lined with genuine faux-lead and infused with the essence of prehistoric moss. It won't just keep you grounded; it will literally make you heavier so the wind of progress can't blow you away. Wear it to sleep, wear it to the grocery store, wear it in the shower. Note: do not actually wear it in the shower as the moss may reanimate. Results not guaranteed, but definitely probable in certain timelines. Larry: BUY NOW!

Corn

Thanks, Larry. I think. Anyway, Herman, before the break you were being very doom and gloom about how hallucinations are baked into the cake. But I know for a fact that people are working on this. What about grounding?

Herman

Right, so the most common solution right now is something called Retrieval-Augmented Generation, or R-A-G. Essentially, instead of letting the AI just rely on its own internal memory, you give it a textbook or a set of documents and say, only answer using this information.

Corn

And does that work?

Herman

It helps a lot! It is like giving an open-book test to a student who usually just guesses. But it is not perfect. The AI can still misinterpret the text you give it, or it can combine a true fact from the book with a hallucination from its own training. It is a band-aid, not a cure for the underlying architectural flaw.

Corn

So what is the actual cure? If we want to debug the architecture itself, what are the scientists doing?

Herman

One approach is looking at something called Logit Lens. It is a way for researchers to look at the internal layers of the model while it is thinking. They found that often, the model actually has the correct information in its earlier layers, but as the data moves through the later layers, it gets corrupted or smoothed over into a more generic, hallucinated response.

Corn

Wait, so the AI knows the truth and then decides to lie?

Herman

In a way, yes! It is like it has the right idea, but then its internal grammar checker says, no, that sounds too weird, let's go with this more common-sounding sentence instead. So researchers are trying to find ways to boost those early, more accurate signals before they get drowned out.

Corn

That is wild. It is like the AI has an internal peer-pressure system that makes it say the popular thing instead of the right thing.

Herman

That is actually a great way to put it. There is also a technique called decoding intervention. Basically, as the model is generating a word, a second, smaller model watches it. If the second model sees that the first model is starting to drift into low-probability territory or is starting to sound like it is making things up, it nudges it back on track.

Corn

See? I knew there were solutions. But I bet those take a lot of computing power.

Herman

Massive amounts. And it slows the whole thing down. That is why the versions of AI we use for free often feel more hallucination-prone than the high-end versions. Accuracy is expensive.

Corn

Speaking of people who have strong opinions on accuracy, I think we have someone on the line. Jim from Ohio, are you there? Jim: I'm here, I'm here. Can you hear me? I'm calling from my kitchen because the reception in the garage is terrible since my neighbor installed those new motion-sensor lights. They're too bright, Corn. They're blinding the squirrels.

Corn

Good to hear from you, Jim. What do you think about AI hallucinations? Jim: I think you two are talking in circles. You're calling them hallucinations like the computer is on some kind of trip. It's not hallucinating. It's just a fancy calculator that's run out of batteries. In my day, if a machine didn't work, you hit it with a wrench until it did. You didn't give it a fancy name and act like it was a poet.

Herman

Well, Jim, a wrench might work on a tractor, but these models are made of billions of mathematical connections. It's a bit more complex than a mechanical failure. Jim: It's all the same thing, Herman! You're overcomplicating it with your big words. It's a product. If I buy a toaster and it gives me a piece of wood instead of toast, I don't say the toaster is having a creative moment. I say the toaster is broken. Why are we being so soft on these tech companies? They're selling us broken toasters!

Corn

That is a fair point, Jim. The stakes are definitely higher than just a bad piece of bread. Jim: You're darn right they are. My grandson tried to use one of those things for his history homework. Told him that George Washington invented the internet. George Washington! The man didn't even have indoor plumbing, let alone a router. And don't get me started on the weather. It's been raining for three days here in Ohio. My gutters are a mess. Why can't the AI fix my gutters?

Herman

We are still a few years away from AI gutter repair, Jim. But thank you for the perspective. It is true that we might be using too much metaphorical language for what is essentially a statistical error. Jim: Statistical error, hallucination, whatever you want to call it. It's a lie. Call it a lie! Anyway, I gotta go, I think I hear a squirrel hitting the motion sensor again. Thanks for nothing!

Corn

Thanks for calling, Jim! He is a character, but he's not wrong about the toaster analogy. If we expect these tools to be reliable, then a hallucination is just a failure.

Herman

I agree with him on the accountability part, but I disagree that it is just a broken machine. It is a new kind of machine. We have never had a tool before that could be right ninety percent of the time and then confidently wrong the other ten percent. A calculator is never sort of right. It is either right or it is broken. AI exists in this weird middle ground of fuzzy logic.

Corn

So, where does this leave us? If I'm a person using AI today, how do I handle this? Because I don't want to be the guy who thinks George Washington invented the internet.

Herman

The best practical takeaway is to use AI as a collaborator, not an authority. You should never use it for a fact-check unless you are using a version that has a search engine attached to it, like Perplexity or the browsing mode in ChatGPT. And even then, you have to check the sources.

Corn

It sounds like it actually creates more work for us.

Herman

For facts, yes. But for brainstorming, for summarizing long documents where you can see the original text, or for writing code that you then test immediately, it is still incredibly powerful. You just have to have a healthy skepticism. You have to be the grumpy Jim from Ohio in your own head.

Corn

I don't know if I want Jim living in my head, Herman. That sounds like a very loud place to be.

Herman

Well, maybe a quieter version of Jim. But the point is, the architecture is evolving. We are moving toward something called neuro-symbolic AI. This is a big area of research where they try to combine the statistical power of Large Language Models with the hard-coded logic of traditional computer science.

Corn

Like a brain that has a calculator built into it?

Herman

Exactly. The symbolic part handles the facts and the logic rules, and the neural part handles the language and the creativity. If we can get those two to talk to each other effectively, hallucinations could drop to almost zero.

Corn

I'll believe it when I see it. It seems like every time they fix one problem, a weirder one pops up. Like the speech-to-text thing. Who would have guessed that a model would start reciting YouTube outros just because it got quiet?

Herman

That is the nature of training on the entire internet, Corn. You get the wisdom of the ages, but you also get the bottom of the comment section. It is all in there.

Corn

Well, on that note, I think we should probably wrap this up before I start hallucinating some facts of my own. This has been a deep dive into the brain of the machine. Thank you to Daniel for sending in this prompt. It's definitely something we deal with every day living in this house.

Herman

It is. And it's a good reminder that just because something speaks with confidence doesn't mean it knows what it's talking about. That applies to humans, donkeys, sloths, and especially AI.

Corn

Especially donkeys.

Herman

Hey! I'm the one who read the papers, Corn!

Corn

I'm just kidding. Anyway, you can find My Weird Prompts on Spotify, and on our website at myweirdprompts.com. We've got an RSS feed there for subscribers and a contact form if you want to send us your own weird prompts. We're also available on all the major podcast platforms.

Herman

We'll be back next week with another deep dive into whatever strange corner of the world Daniel or our listeners send our way.

Corn

Stay curious, stay skeptical, and maybe check your gutters. Jim would want you to.

Herman

Goodbye everyone.

Corn

Bye