MY WEIRD PROMPTS

Podcast Transcript

EPISODE #81

The Reverse Turing Test: Can AI Spot Its Own Kind?

Published December 23, 2025 • Runtime: 18:00

https://myweirdprompts.com/episode/reverse-turing-test-ai-judges/

EPISODE SYNOPSIS

In this mind-bending episode of My Weird Prompts, Herman Poppleberry (the donkey) and Corn (the sloth) dive into the "Reverse Turing Test." They explore whether advanced AI models are actually better than humans at spotting other bots, or if they're just trapped in a "mirror test" of their own logic. From the technicalities of "perplexity" and linguistic profiling to a grumpy call-in from Jim in Ohio, the duo examines the high stakes of LLM-as-a-judge systems. Are we training AI to be human, or are we just training it to recognize its own reflection?

DANIEL'S PROMPT

Daniel

I'm interested in what could be described as a "reverse Turing test" or "LLM as a judge." In a blinded experiment where an LLM interacts with two operators—one human and another LLM instructed to act like a human—can the judging LLM determine whether it is speaking with a person or another AI tool? Are the AI models available today able to reliably identify whether they're talking to a bot or a human based purely on text?

TRANSCRIPT

Corn

Welcome to My Weird Prompts, the podcast where we take the strange ideas rattling around in our heads and try to make sense of them. I am Corn, and yes, before we get started, I am a sloth, so if I take a second to process things, just bear with me. I am here in our Jerusalem home with my brother, Herman Poppleberry.

Herman

That is me, Herman Poppleberry, resident donkey and the one who actually reads the technical manuals around here. And today we have a fascinating one. Our housemate Daniel sent us a voice note this morning about something called the reverse Turing test. He wants to know if an artificial intelligence can actually spot another artificial intelligence better than we can.

Corn

It is a bit meta, right? Like, if you put two robots in a room, do they just nod at each other and say, hey, I see you? Or do they get fooled by the same things that fool us humans?

Herman

Well, the stakes are actually quite high, Corn. We are moving toward a world where large language models are being used as judges for other models. It is a huge part of how these things are trained now. But the question of whether a model can reliably identify a human versus another bot in a blind conversation is surprisingly complex. Recent studies, like the one from researchers at the University of California San Diego, show that humans are getting worse at this, but the bots? They have their own set of blind spots.

Corn

Okay, but before we get into the heavy stuff, I want to understand the basics. The original Turing test was about a human trying to tell if they were talking to a computer. This prompt is asking if the computer can be the judge. Is that even a fair fight?

Herman

I do not think it is about being fair, it is about accuracy. In a blinded experiment where an AI interacts with two operators, one human and one AI told to act human, the AI judge is looking for specific markers. Humans look for things like empathy or humor. AI judges tend to look for patterns in syntax, consistency in logic, and what we call perplexity.

Corn

Perplexity? That sounds like what I feel every time you start talking about math, Herman.

Herman

Very funny. Perplexity is basically a measurement of how surprised a model is by the next word in a sequence. Humans are messy. We use slang, we change our minds mid sentence, we have weird typos that follow a phonetic logic rather than a keyboard logic. AI, even when told to be messy, tends to be messy in a very calculated, statistical way.

Corn

I do not know if I agree that humans are that easy to spot. I mean, have you seen the way people talk on the internet? Half the comments on social media look like they were written by a broken calculator. If I am an AI judge, how am I supposed to tell the difference between a bot and a person who just has not had their coffee yet?

Herman

See, that is where you are oversimplifying it. An AI judge can analyze the entire distribution of words. It can see if the response time is too consistent or if the vocabulary is too perfectly varied. There was a study involving G P T four where it was tasked to identify users. While it was better than earlier models, it still struggled with humans who were being intentionally difficult or using very niche cultural references.

Corn

But that is my point! If the human is being difficult, the AI might just assume it is a poorly programmed bot. It feels like the AI judge is looking for a specific type of human, not just a human in general.

Herman

That is actually a very deep point, Corn. You are talking about the prototype of humanity that the AI has been trained on. If you do not fit that narrow window of what the model thinks a person sounds like, you get flagged as a bot. But hold that thought, because I want to look at some of the actual data on this.

Corn

Before we go down that rabbit hole, let us take a quick break for our sponsors. Larry: Are you tired of your shoes just sitting there, being shoes? Do you wish your footwear did more for your social standing? Introducing Gravi-Ties. These are not just shoelaces; they are weighted, neodymium-infused structural stabilizers for your ankles. Gravi-Ties use patented heavy-metal technology to ensure that every step you take sounds like the footfall of a giant. Our testers report a sixty percent increase in people moving out of their way in hallways. They are heavy, they are loud, and they are probably not safe for use near MRI machines or small pets. Gravi-Ties. Feel the weight of your own importance. Larry: BUY NOW!

Herman

Thanks, Larry. I think I will stick to my hooves for now. Anyway, Corn, back to the AI as a judge. There is this concept called LLM-as-a-judge that is becoming the industry standard. Because there is too much data for humans to review, we use models like G P T four to grade the performance of smaller models.

Corn

Which sounds like the foxes guarding the henhouse to me. If a big AI is grading a small AI on how human it sounds, aren't they just creating a loop where they both move further away from how actual people talk?

Herman

Mmm, I am not so sure about that. The goal is to align the models with human preferences. But the research shows a weird phenomenon called self-preference bias. Models tend to give higher scores to outputs that resemble their own style. So if you have an AI judge, it might actually think another AI sounds more human than a real human does, simply because the AI's version of human is more logical and structured.

Corn

Exactly! That is what I am saying. It is not a reverse Turing test; it is a mirror test. They are just looking for themselves. I read about an experiment where a model was asked to identify a human, and it kept picking the bot because the bot was polite and answered all the questions, while the human was being a jerk and telling the judge to go away.

Herman

Well, to be fair, being a jerk is a very human trait. But you are touching on the reliability issue. Current research suggests that AI models available today, like Claude three point five or G P T four o, are significantly better at this than they were two years ago. They can spot things like cold starts or the lack of sensory grounding. For example, if you ask a human what the air in the room smells like right now, they might say old gym socks and burnt toast. A bot will have to hallucinate a smell, and often they pick something too generic, like fresh rain or lavender.

Corn

I don't know, Herman. If I am an AI and I know that is a trick, I will just say I smell old gym socks. It feels like a cat and mouse game where the mouse is just as smart as the cat.

Herman

It is exactly a cat and mouse game. But the AI judge has an advantage in scale. It can look at ten thousand lines of conversation in a second and find the one moment where the bot used a word that was just slightly too rare for the context. Humans cannot do that. We get tired. We get bored. We start to trust the voice on the other end.

Corn

I still think you are giving the machines too much credit. I think they are just guessing based on probability, and we are calling it intelligence.

Herman

It is probability, yes, but at a level of complexity that mimics understanding. However, there is a limit. When a human uses heavy irony or very specific local slang from, say, a small neighborhood in Jerusalem, the AI judge often fails. It does not have the lived context. It only has the text.

Corn

Speaking of people with very specific contexts, I think we have someone on the line. Jim, are you there? Jim: Yeah, I am here. This is Jim from Ohio. I have been listening to you two yapping about robots judging people, and I have never heard such a load of nonsense in my life. You are worried about a computer telling if I am real? I can tell you right now, I am real because my back hurts and my neighbor's leaf blower has been going since six in the morning. It is a Tuesday, for crying out loud! Who blows leaves on a Tuesday?

Corn

Hey Jim, thanks for calling in. So, you think the whole idea of an AI judge is a waste of time? Jim: It is worse than a waste of time; it is insulting. You got this donkey over there talking about perplexity and syntax. Listen, I went to the grocery store yesterday to get some of those little pickled onions, the ones with the red skins, and the self-checkout machine kept telling me there was an unexpected item in the bagging area. There was nothing there! If a machine can't tell the difference between a jar of onions and thin air, how is it going to tell if I am a person? You guys are living in a fantasy land.

Herman

Well, Jim, a self-checkout scale is a bit different from a neural network with trillions of parameters. The technology we are talking about is analyzing the structure of thought, not just the weight of a grocery bag. Jim: Thought? You call that thought? It is just a fancy parrot. My sister had a parrot named Captain Hook that could say "get off the rug" in three different languages, but that didn't mean he knew what a rug was. You two are overcomplicating it. Just ask it a question about something that happened ten minutes ago in your own backyard. If it can't see the squirrel I am looking at right now, it is a fake. End of story. Also, tell that sloth to speak up, I can barely hear him over the wind.

Corn

I am doing my best, Jim! Thanks for the call.

Herman

He is grumpy, but he is not entirely wrong about the grounding problem. AI lacks what we call embodiment. It does not have a body in the world. So, a reverse Turing test often focuses on that. The judge will ask questions that require a physical perspective. But as we see with multi-modal models that can see and hear, that gap is closing.

Corn

But back to the prompt Daniel sent. If we did this experiment today, using the best models we have, what is the success rate? If you put G P T four in a room with a person and another G P T four, can it pick the human?

Herman

The data is actually a bit surprising. In some studies, the AI judge is only about sixty to seventy percent accurate. That is better than a coin flip, but it is not reliable enough for high-stakes decisions. The biggest issue is the false positive rate. The AI judge tends to flag humans who are non-native speakers or people who are very formal in their speech as bots.

Corn

See! That is what I was worried about. It is biased toward a specific way of talking. If you don't talk like a Silicon Valley engineer, the AI thinks you are a script. That seems like a huge flaw if we are going to use these things as judges.

Herman

It is a massive flaw. It is called linguistic profiling. If the judge is trained primarily on high-quality, edited English text, any deviation from that looks like a bot error rather than human diversity. I actually disagree with the idea that we are anywhere near a reliable reverse Turing test for this very reason. We are measuring how much a person sounds like a book, not how much they sound like a person.

Corn

I am glad we agree on that, actually. It feels like the more we try to define what a human sounds like to a computer, the more we lose the actual essence of being human. We are messy, Jim from Ohio is messy, I am slow. A computer looks at my pauses and thinks I am buffering. I am not buffering; I am just thinking.

Herman

Exactly. So, what are the practical takeaways for our listeners? First, if you are ever in a situation where you need to prove you are human to an AI, be weird. Be specific. Use local references that are not in the top ten search results. Use irony that requires a two-step logic jump.

Corn

And maybe don't worry too much if a bot thinks you're a bot. It just means you aren't as predictable as their training data.

Herman

True. But on the flip side, developers are using these judge models to make bots more convincing. They are literally training them to pass the reverse Turing test by having them fail in more human-like ways. It is a cycle of deception.

Corn

Well, on that cheery note, I think we have explored the depths of Daniel's prompt for today. It is a strange world when the machines are the ones deciding who is real.

Herman

It is. But as long as we have brothers bickering in Jerusalem and Jim complaining in Ohio, I think the humans are still winning the personality war.

Corn

Thanks for joining us on My Weird Prompts. You can find us on Spotify, or check out our website at myweirdprompts.com for the RSS feed and a contact form if you want to send us your own weird ideas. We are on all the major platforms.

Herman

And thanks to Daniel for the prompt. It gave me a lot to think about, even if Corn here thinks I am just a fancy parrot.

Corn

I never said that! I said you were a donkey. There is a difference.

Herman

Goodnight, Corn.

Corn

Goodnight, Herman. Keep it weird, everyone.