Podcast Transcript

EPISODE #46

Pixels, Prompts & Pseudo-Text: AI's Word Problem

Published December 10, 2025 • Runtime: 23:53

https://myweirdprompts.com/episode/pseudotext/

EPISODE SYNOPSIS

Why can advanced AI models generate breathtaking photorealistic landscapes and fantastical creatures with astonishing detail, yet consistently stumble over spelling a simple word like 'cat' on a t-shirt? This week on My Weird Prompts, co-hosts Corn and Herman dive into producer Daniel Rosehill's intriguing prompt: the pervasive and often comical challenge of 'pseudo-text' in AI image generation. They unpack the fundamental distinction between how AI processes visual information at a pixel level versus its understanding of symbolic language, revealing why generating coherent text within images is a far more complex multi-modal problem than it appears. Explore the cutting-edge "pipelined" solutions that integrate language models to improve accuracy, and

TRANSCRIPT

Corn

Welcome, welcome, welcome to "My Weird Prompts"! I'm Corn, your ever-curious co-host, and as always, I'm joined by the encyclopedic Herman.

Herman

Greetings, Corn, and to all our discerning listeners tuning in from across the digital airwaves.

Corn

So, Herman, this week's prompt comes from none other than our show's producer, Daniel Rosehill. And it's a fascinating deep dive into something I think a lot of our listeners, especially those dabbling in AI art, have probably stumbled upon.

Herman

Indeed. Daniel's prompt zeroes in on what he calls "pseudo-text" in AI image generation. It's a persistent, almost comical, challenge in the otherwise astonishing progress of these models. Most people marvel at the photorealistic landscapes or fantastical creatures AI can conjure, but few realize it often struggles with something as fundamental as spelling "cat" correctly on a t-shirt.

Corn

Right! It feels almost counter-intuitive, doesn't it? Like, you can generate a hyper-realistic image of a sloth wearing a tiny hat, but if you ask that hat to say "Hello," it comes out looking like ancient alien script. What is *that* about?

Herman

Well, that's exactly the core of Daniel's question, and it highlights a fundamental distinction in how these models perceive and generate information. We're going to explore why image generation has advanced so dramatically, yet text generation *within* those images remains such a stubborn hurdle. It's not just a minor glitch; it reveals a lot about the current limitations and future directions of AI.

Corn

And I think it's important because so many of us are using these tools now. We see these incredible images online, and then we try to make something with text, and it's like the AI suddenly forgets how to read or write. It's almost like it's mocking us with these garbled letters.

Herman

"Mocking" might be a strong word, Corn, but I understand the sentiment. It's certainly frustrating when you're expecting a coherent message and you get what looks like a randomized selection of glyphs from a forgotten language.

Corn

So, Herman, you mentioned a fundamental distinction. Lay it on me. Why can an AI create a breathtaking vista, but trips over "OPEN" on a storefront sign?

Herman

Excellent question. At its heart, generative AI for images, especially large diffusion models, operates on a pixel level. It's learning patterns of light, color, and form. When it generates an image, it's essentially predicting what pixels should go where to create a coherent visual scene based on its training data. It's a probabilistic exercise in visual composition.

Corn

Okay, so it sees "tree" and it knows what a tree generally looks like, where leaves go, how the bark looks. It's painting with pixels.

Herman

Precisely. Now, consider text. When you see the word "tree," you don't just see a collection of pixels. You recognize specific shapes – the letters T, R, E, E – arranged in a particular sequence. More than that, you understand that sequence carries semantic meaning. It's a symbolic representation.

Corn

Ah, so it's not just about drawing the *shape* of a letter 'T', it's about understanding that 'T' is a specific character in an alphabet, and it combines with other specific characters to form a word that means something.

Herman

Exactly. For an AI model trained primarily on images, text appears as just another visual texture or pattern. It learns to approximate the *appearance* of text – the blocky shapes, the lines, the way letters usually group together – but it doesn't understand the underlying symbolic structure or the grammar of language. It's like a highly skilled artist who can perfectly copy a foreign alphabet's script, but doesn't speak a word of the language. They reproduce the visual form, but not the meaning.

Corn

That's a great analogy! So when Daniel tells the AI to put "It's fun to be a sloth" on a t-shirt, the AI isn't *reading* those words. It's essentially trying to *draw* what it thinks words look like based on all the images of text it's seen.

Herman

That's a fair way to put it. It might pick up on common letter shapes, the density of characters within a word, or even the general "word-like" patterns. But without a deep linguistic understanding, it struggles with the precise ordering, the correct letter forms, and the contextual meaning. This is why you often get garbled letters, transposed characters, or entirely random symbols that vaguely resemble text.

Corn

So, you're saying it's not just about resolution or the fidelity of the image, but a completely different cognitive process for the AI?

Herman

In essence, yes. Image models excel at synthesizing visual information. Language models excel at synthesizing symbolic, linguistic information. Marrying the two, where the image model correctly *generates* the visual form of specific, semantically correct text *within* an image, is a far more complex challenge than it appears. It requires the AI to operate effectively in two very different domains simultaneously. It's a multi-modal problem, and frankly, the current diffusion architectures aren't inherently optimized for this kind of precise symbolic rendering within a visual field.

Corn

But wait, Herman. If that's the case, then why are some models getting better? Daniel mentioned Gemini, and how it's getting closer to "90% plus accuracy" for tasks like "whiteboard cleaning," where you give it a sketch and it turns it into a flowchart. That sounds like it's understanding text, or at least converting it.

Herman

You're jumping ahead a bit, Corn. While Daniel's observations about Gemini's performance on tasks like "whiteboard cleaning" are valid and exciting, it's important to understand the *nature* of that improvement. It's not necessarily that the image generation model itself has suddenly become a linguistic expert.

Corn

Okay, so what is it then? Are you saying it's a trick?

Herman

Not a trick, but a different approach. Modern multi-modal models like Gemini often incorporate language models (LLMs) *alongside* their image generation components. So, when you ask it to "clean a whiteboard" and convert handwritten text to a flowchart, what's likely happening is that the image model first performs an advanced form of Optical Character Recognition, or OCR, to *extract* the text. That extracted text is then processed by a language model, which can understand the semantic content, correct spelling, and structure it. Finally, the image generation part *re-renders* the new, corrected text back into the image, often using predefined font styles or generating highly legible approximations.

Corn

So, it's less about the AI *drawing* the text perfectly from scratch, and more about it being a pipeline of different AI systems working together?

Herman

Exactly. It's a more sophisticated workflow. The image generation portion is still excellent at *incorporating* legible text once it's been correctly generated by a separate, dedicated language component. The challenge of a single, unified diffusion model flawlessly *creating* accurate text from a purely visual understanding remains significant.

Corn

That makes a lot more sense. It's like having a skilled transcriptionist working with a graphic designer, rather than asking the graphic designer to decipher squiggles and then perfectly draw out the correct words without understanding them.

Herman

A very apt analogy, Corn. And this "pipelined" approach is why we're seeing improvements in those specific use cases. The AI isn't just "trying its best" to draw text; it's leveraging its language understanding where it can.

Corn

Let's take a quick break from our sponsors. Larry: Are you feeling overwhelmed by the sheer volume of information in today's world? Do you struggle to make sense of complex ideas, or even basic instructions? Introducing "Clarity Crystal"! This revolutionary, pocket-sized device uses proprietary vibrational frequencies to instantly distill any confusing information into crystal-clear understanding. Just hold it near a book, a screen, or even a mumbled conversation, and feel the knowledge osmose directly into your brain! No more overthinking, no more deciphering pseudo-text! Clarity Crystal: because why read when you can just... *know*? Limited supplies, no refunds. BUY NOW!

Herman

...Alright, thanks Larry. Anyway, Corn, picking up where we left off, even with these pipeline approaches, Daniel mentioned those "little pseudo-text problems" still pop up. What's going on there? If it's using an LLM to generate the text, shouldn't it be perfect?

Corn

Yeah, I mean, 90% accuracy is good, but that remaining 10% can be really frustrating, especially if it's a crucial piece of text. What causes those lingering errors?

Herman

Well, the pseudo-text issues can still arise from several points in that pipeline. First, the initial OCR component might not be 100% accurate, especially with highly stylized fonts, unusual layouts, or very messy handwriting. It's not foolproof.

Corn

So, it's still possible for it to misread the original handwritten text, even if it's supposed to be "cleaning" it up?

Herman

Absolutely. Think about how difficult it can be for *humans* to read certain types of handwriting. The AI, while advanced, isn't immune to those challenges. Second, even if the text is correctly extracted and processed by the LLM, the final image generation step, where that text is rendered back into the image, can still introduce artifacts. This is where the old "drawing" problem reappears. The AI might slightly distort a letter's shape, misalign characters, or struggle with intricate kerning – the spacing between letters – especially if it's trying to integrate the text seamlessly into a complex visual scene.

Corn

So, it's like a game of telephone, where information can get corrupted at different stages. The OCR might mishear, the LLM might interpret, and the final image generator might misdraw.

Herman

A reasonable analogy. Furthermore, the *style* of text is another layer of complexity. If you want a specific font, a certain texture, or even a distressed look for the text, that requires the image generation component to have a very fine-grained control over visual details, which can sometimes conflict with maintaining perfect legibility. It's a trade-off.

Corn

That's a good point. I've seen AI-generated images where the text looks "melted" or like it's part of the background texture, which is cool for aesthetics, but terrible if you actually need to read it.

Herman

Exactly. And Daniel also noted that specialized models designed specifically for text replication haven't necessarily offered better results than general models in practice. This is often because specialized models, while theoretically precise, might lack the broad understanding and versatility of the larger, general-purpose models. They can sometimes be overfit to specific text generation tasks and fail when presented with slightly different styles or contexts.

Corn

So, bigger models, even with their quirks, are still more robust overall?

Herman

In many cases, yes. The sheer scale and diversity of their training data give them a wider range of capabilities, even if they're not perfectly optimized for every single sub-task. It's a classic breadth versus depth dilemma.

Corn

And we've got a caller on the line. Go ahead, you're on the air. Jim: Yeah, this is Jim from Ohio. And I've been listening to you two go on and on about this "pseudo-text" and frankly, it just sounds like bad programming. My old VCR used to put the time on the screen and it never messed up the numbers. You guys are making a mountain out of a molehill. Also, my neighbor Gary just bought one of those fancy electric cars, and it's so quiet I almost ran him over this morning backing out of my driveway. Nearly spilled my coffee too. But seriously, just make the computer spell it right. It's not that hard.

Herman

Jim, I appreciate your perspective, and I understand why it might seem straightforward. However, the VCR example is a bit different. That VCR had a very limited, predefined set of characters and positions it could display. It wasn't *generating* new images with complex text integrated into a scene. It was displaying pre-programmed graphics.

Corn

Yeah, Jim, think of it this way: the AI isn't just putting a sticker on an image. It's trying to *paint* the letters perfectly into a dynamic, often irregular surface, while also making sure they fit the style and context of the image. It's much more complex than just displaying a font. Jim: Nuance, shmoonance. If these machines are so smart, they should be able to spell. My cat Whiskers can almost spell her name with her paw, and she's just a cat. All this talk of "pixel levels" and "linguistic models" just sounds like excuses. Back in my day, if something didn't work, you fixed it, simple as that. Not all this highfalutin explanation.

Herman

Jim, it's not an excuse; it's an explanation of an ongoing technical challenge at the forefront of AI research. We're talking about systems that are still learning to bridge the gap between visual representation and symbolic meaning. It's a fascinating problem, actually.

Corn

It really is, Jim. And while we appreciate your frankness, it's not quite as simple as just "fixing it" when you're dealing with emergent behaviors in neural networks. But thank you for calling in! Jim: Eh, whatever. You guys are still missing the point. My knee's been acting up today anyway, probably from that rain we had. Good day.

Corn

Thanks for calling in, Jim! Always good to hear from Ohio. Herman, Jim's got a point in a way – from a user perspective, if the AI can do all these amazing things, why *can't* it spell? What can listeners actually do with this information now? How do we navigate this pseudo-text problem?

Herman

It's a valid user frustration, Corn, and the practical takeaways are important. First, for critical text that absolutely *must* be accurate and legible within an AI-generated image, the safest approach for now is often a hybrid one. Generate the image without the text, then use traditional graphic design tools or even other AI text-to-image tools that specifically excel at typography to *overlay* or *insert* the text afterwards.

Corn

So, use the AI for the pretty picture, and then bring in Photoshop or Canva for the text. That makes sense for high-stakes stuff.

Herman

Precisely. Second, when you *do* try to generate text directly within an AI image model, try to keep the text short, simple, and in a common, highly legible font. Avoid overly complex fonts, unusual letter combinations, or very long phrases. The less ambiguity, the better.

Corn

So, "STOP" is better than "Stop in the name of love"?

Herman

Significantly better. And third, experiment with different prompts and models. As Daniel noted, some models are making progress. Using specific phrasing like "clear, legible text," "sans-serif font," or "text on a plain, contrasting background" can sometimes guide the AI more effectively, though results will still vary.

Corn

And I'd add, manage your expectations. If you know there's a strong chance the text will be garbled, either have a backup plan or embrace the weirdness! Some of the pseudo-text failures can actually be quite artistic or funny.

Herman

I'd push back on "embrace the weirdness" for professional applications, Corn. For personal fun, perhaps. But for anything requiring clear communication, accuracy is paramount. It’s better to use the right tool for the job.

Corn

Fair enough. But for creators just experimenting, the unexpected can be inspiring, sometimes! Herman, what do you see for the future? Will AI eventually just nail text generation within images, or will this always be a bit of a tricky dance?

Herman

I believe we will eventually see significant breakthroughs. The current trajectory points towards increasingly sophisticated multi-modal architectures that can seamlessly integrate linguistic and visual understanding. We're already seeing models that can edit specific objects or styles within an image with remarkable precision; applying that same precision to symbolic text is the next logical step.

Corn

So, no more weird squiggles that look like ancient alien messages?

Herman

Ideally, no. We'll likely see advancements in what are called "compositional" AI models that have a more explicit understanding of how different elements, including text, should be structured and interrelated within a visual scene. This could involve new training methodologies or even novel architectural designs that specifically address the gap between pixel-level rendering and symbolic meaning. It's a hard problem, but not an insurmountable one.

Corn

That's a hopeful thought. It really illustrates how despite all the incredible progress, AI is still very much a field of active research, with fundamental challenges like this pseudo-text problem still pushing the boundaries.

Herman

Precisely. It keeps us on our toes, doesn't it? And it reminds us that while AI can perform astonishing feats, it often struggles with what we humans consider trivial, like simply spelling a word.

Corn

Absolutely. It makes you appreciate the human brain's ability to effortlessly blend language and visual perception. A truly weird prompt this week from Daniel Rosehill, sparking some great discussion.

Herman

Indeed. A thought-provoking prompt, as always.

Corn

And that's all the time we have for "My Weird Prompts" this week. A huge thank you to Herman for breaking down the complexities of AI and pseudo-text.

Herman

My pleasure, Corn. Always a stimulating conversation.

Corn

You can find "My Weird Prompts" on Spotify and wherever else you get your podcasts. Make sure to subscribe so you don't miss an episode. Until next time, keep those weird prompts coming!