MY WEIRD PROMPTS

Podcast Transcript

EPISODE #23

AI's Blind Spot: Data, Bias & Common Crawl

Published December 05, 2025 • Runtime: 34:41

https://myweirdprompts.com/episode/ais-blind-spot-data-bias-common-crawl/

EPISODE SYNOPSIS

In this eye-opening episode of "My Weird Prompts," hosts Corn and Herman dive deep into the unseen influences shaping large language models. They explore the critical topic of AI training data, uncove...

DANIEL'S PROMPT (Summary)

Daniel

Episode from My Weird Prompts podcast

TRANSCRIPT

Corn

Welcome to "My Weird Prompts"! I’m Corn, your endlessly curious host, and I’m joined as always by the encyclopedic Herman. Daniel Rosehill, our producer and the brilliant mind behind this podcast, has really sent us down a rabbit hole this week.

Herman

Indeed, Corn. Daniel wanted us to explore a topic that’s fundamental to the very existence and behavior of large language models: their training data. But not just *what* it is, but *where* it comes from, the biases it instills, and some profound questions around consent and transparency.

Corn

It sounds like we're not just looking at the nuts and bolts of how AI learns, but also peeling back the layers to understand the unseen influences shaping these powerful tools. And Daniel specifically mentioned a "blind spot" in the discussion around AI censorship and bias. Herman, can you kick us off by explaining what Daniel means by that?

Herman

Certainly, Corn. Daniel highlights a crucial distinction. We often think of censorship as overt suppression, like what he mentioned regarding models from certain regions like Asia, where state-imposed guardrails are explicit. However, in other contexts, particularly what he terms "American-centric" environments, the issue isn't overt censorship but rather a more subtle, inherent bias baked into the model's worldview.

Corn

So, it's not someone actively saying "don't talk about this," but more like the model simply doesn't have the perspective or information to discuss it in a balanced way, or it defaults to a particular cultural viewpoint.

Herman

Precisely. It's a byproduct of the training data. If the vast majority of information an AI model consumes reflects a predominantly American cultural perspective, then its responses, its understanding of societal norms, its examples, and even its humor, will naturally lean in that direction. This isn't a malicious design; it's an emergent property of the data it has learned from. For instance, if you ask about common holidays, it might prioritize Thanksgiving or the Fourth of July, not because it's programmed to, but because its training data had a much higher frequency of discussions about those specific cultural events.

Corn

That makes a lot of sense. I can see how that would degrade the "authenticity" of the experience, as Daniel put it, if you’re, say, in Jerusalem, and the AI is constantly making references that don’t quite land with your local context. It's like trying to have a conversation with someone who only watches American sitcoms and then uses those references to explain the world.

Herman

A very apt analogy, Corn. And Daniel’s prompt raises the question of *why* this happens. The answer, as he alluded to, lies in the composition of the training data. He mentioned colossal collections of textual information from sources like Reddit, Quora, and YouTube. These platforms, while global, have a very strong Western, often American, user base and content output, especially in the English language. This creates a data imbalance.

Corn

Okay, so it’s not just the sheer volume of data, but the *origin* and *nature* of that data. I mean, Reddit is a huge forum for discussion, but it has its own subcultures, its own biases, its own humor. The same for Quora and YouTube comments. If these models are feasting on that, they're absorbing all of that, too.

Herman

Absolutely. Think of it this way: these large language models, or LLMs, aren't truly intelligent in the human sense. They are, at their core, sophisticated pattern recognition and prediction machines. They predict the next most probable token or word based on the vast dataset they've consumed. If that dataset is heavily skewed towards certain cultural expressions or viewpoints, the model's output will reflect that statistical bias. It's a mirror, albeit a very complex one, reflecting the internet's dominant voices back at us.

Corn

So if I'm, say, looking for a nuanced political perspective on a non-Western country, and the training data is primarily from Western sources, the AI's response might inadvertently frame it through a Western lens, even if it tries to be objective. That's a huge implication.

Herman

Indeed. And it speaks to the challenge of creating truly universal AI. To address this, developers would need to curate incredibly diverse, globally representative datasets, which is an immense undertaking, both technically and logistically.

Corn

Speaking of immense undertakings, Daniel also brought up something called "Common Crawl." He even joked that before the LLM era, its description sounded "slightly shady." Herman, what is Common Crawl, and why is it so significant in this discussion?

Herman

Common Crawl, Corn, is a non-profit organization founded in 2007. Its official description, which Daniel read, states they "make colossal scale extraction, transformation, and analysis of open web data accessible to researchers." In simpler terms, they crawl the web, much like a search engine, but instead of just indexing pages, they collect the raw data from billions of web pages. This massive archive of web data is then made publicly available as datasets.

Corn

So, they're essentially creating a gigantic snapshot of the internet over time? That's… well, it certainly sounds less shady now that I understand the intent, but the scale is mind-boggling. Daniel mentioned their latest crawl is "freely accessible," bundling "the entire internet." Is that an exaggeration?

Herman

Not significantly. While "the entire internet" is a poetic overstatement—the internet is constantly evolving and vast beyond any single capture—Common Crawl's datasets are indeed enormous. The latest crawl, CC-MAIN-2023-47 for instance, encompasses petabytes of data, containing billions of web pages. It's distributed in formats like WARC, WAT, and WET files, which are essentially archives of web content designed for large-scale processing. So, while you can download a segment of it, downloading and processing the *entire* crawl requires significant computational resources, far beyond what a typical individual could manage on a home computer.

Corn

So, it's not something I can just casually put on a USB stick, like Daniel mentioned. This is industrial-scale data. And it sounds like this is the prime feeding ground for many of these large language models we interact with today.

Herman

Absolutely. For anyone looking to train an LLM from scratch, particularly those outside the absolute top-tier tech giants with their own proprietary web-crawling infrastructure, Common Crawl is an invaluable, foundational resource. It provides a vast, relatively clean, and publicly available dataset of human language and information from the web. It significantly lowers the barrier to entry for researchers and developers to build and test their own models.

Corn

And that leads directly to Daniel's concern about the "consent" aspect. He talked about people posting on Reddit in 2008, or writing a blog about Jerusalem, long before large language models were even a household name. They weren't thinking, "Oh, this is going to be scraped by an AI to train its brain." How do we reconcile that?

Herman

This is one of the most significant ethical dilemmas surrounding LLM training data, Corn. When Common Crawl started, and when much of its historical data was collected, the concept of AI models ingesting and learning from vast swathes of the internet was largely confined to academic research or speculative science fiction. Users posting online were implicitly agreeing to terms of service for *that specific platform*, not explicitly consenting to their data being used to train generative AI.

Corn

So, the "consent" we implicitly gave to, say, Facebook or Reddit, wasn't for this new, unforeseen use case. And now, that data has been "swallowed up," as Daniel put it, by these bundling projects.

Herman

Precisely. Common Crawl does have an opt-out registry, where website owners can request their content not be crawled. However, as Daniel points out, this is a reactive measure. For historical data, or for individuals who weren't website owners but content creators on platforms, the notion of "retroactive consent" is practically non-existent or, at best, highly problematic. It raises fundamental questions about data ownership, intellectual property rights, and the future implications of publishing anything online.

Corn

I mean, if you wrote a novel on a blog back in 2009, and now bits of it are showing up in an AI's creative writing, without any attribution or compensation, that feels like a violation. Even if it's transformed, the source material is still there.

Herman

It's a complex legal and ethical landscape, Corn. Copyright law is still grappling with how to apply to AI-generated content and its training data. Some argue that because the AI doesn't directly copy but rather learns patterns, it's akin to a human reading and being influenced by texts. Others contend that mass ingestion without explicit consent or licensing constitutes infringement, especially if the original work is recognizable or derivative works impact the market for the original. This is an ongoing legal battle, with several high-profile lawsuits currently making their way through the courts.

Corn

So, for most people posting anything online, it's essentially out there for the taking, even if they didn't realize it at the time. That's a bit unsettling.

Herman

It is. And it underscores a shift in our understanding of digital privacy and public information. What was once considered "public" in a human-readable, human-consumable sense, is now also "public" in a machine-readable, machine-consumable sense, with far-reaching implications that we are only just beginning to grasp.

Corn

Daniel also touched on the fact that LLMs have "training cut-offs." They're not always up-to-the-minute with current events, unless they use "external tooling." Can you explain the difference and why that's important?

Herman

Absolutely. Early LLMs, and even the base models today, are trained on a static dataset up to a specific point in time. This creates a "knowledge cut-off." For example, a model trained up to mid-2023 won't inherently know about events that occurred in late 2023 or 2024. Its "worldview" is frozen at that point.

Corn

So if I asked it about the latest Oscar winners, and its cut-off was last year, it wouldn't know.

Herman

Exactly. To overcome this limitation, developers have integrated what Daniel calls "external tooling." This primarily involves allowing the LLM to access real-time information sources, such as search engines, through an API. When you ask a question about current events, the LLM doesn't "know" the answer from its internal training; instead, it uses a tool to search the web, process the results, and then formulate a response based on that up-to-date information.

Corn

Ah, so it's like giving the AI a web browser and teaching it how to use it. It’s not remembering the new information, it’s just accessing it. That makes sense. It bridges that knowledge gap.

Herman

And this integration is becoming increasingly sophisticated, with models capable of complex multi-step reasoning using various tools. However, the core training data remains foundational, dictating the model's language patterns, general knowledge, and reasoning capabilities, even if the specific facts come from an external source.

Corn

Daniel also highlighted the immense resources required to train an LLM. He called it a "multi-million dollar process, minimum, really hundreds of millions." Why is it so incredibly expensive and out of reach for most individuals?

Herman

The cost stems from several factors, Corn. Firstly, the sheer computational power needed. Training an LLM involves processing petabytes of data, requiring thousands of high-performance GPUs running continuously for weeks or even months. This translates to enormous electricity bills and significant hardware investment. Secondly, data acquisition and curation. While Common Crawl provides raw data, preparing it for training—cleaning, filtering, de-duplicating, and formatting—is a labor-intensive process that often requires human oversight and specialized tools. Thirdly, expertise. Developing and refining these models requires a team of highly skilled machine learning engineers, data scientists, and researchers, whose salaries are substantial.

Corn

So, it's not just the computer time, it's the people time, and the infrastructure to support it all. That explains why an individual like Daniel can't just "create from scratch my large language model," as he put it.

Herman

Precisely. Fine-tuning an existing model is accessible to more people, as it requires less data and computational power. But building a foundational model from the ground up, with billions or trillions of parameters, is an undertaking reserved for well-funded organizations. This concentration of power in the hands of a few large companies also has implications for the diversity of AI models and the potential for new forms of bias.

Corn

Wow. This all ties into the listener question Daniel mentioned: "Is the model being trained on your data?" He talked about how OpenAI states they don't train on your data if you're a paid customer, and he believes them. Can you elaborate on why that's so important and why it's plausible?

Herman

This is a critical point for user trust and data privacy. For paid users, companies like OpenAI often promise not to use their conversational data for training. Daniel’s reasoning for believing this is sound. Firstly, the data generated by individual user prompts and AI responses is highly idiosyncratic. It’s a mix of conversational styles, specific queries, and AI-generated text, which can create a very "noisy" dataset. Trying to extract meaningful, generalizable patterns from this jumbled, uncurated stream of data from millions of users would be incredibly inefficient and difficult to integrate into a foundational model.

Corn

So, instead of being helpful, it would actually make the model worse or at least much harder to train effectively. It's like trying to learn proper grammar by only listening to people having disorganized arguments.

Herman

An excellent analogy. Furthermore, there's the monumental logistical challenge of processing and incorporating that real-time, unstructured user data without introducing biases or, more critically, privacy violations. The sheer volume and diversity of user interactions would make quality control a nightmare. The second, and perhaps more compelling reason, as Daniel noted, is the potential for scandal.

Corn

Ah, the "penny drop" scenario. If proprietary company secrets or personal medical information from a user’s prompt started showing up in the AI's public responses to *other* users, that would be a catastrophic breach of trust and likely lead to massive lawsuits.

Herman

Exactly. The reputational and legal damage would be immense. Given the stakes, it's far safer and more practical for these companies to rely on massive, pre-curated datasets like Common Crawl for foundational training, and then potentially fine-tune models on internal, anonymized, and aggregated user data under strict controls, if at all. Daniel's personal choice to avoid free models that *do* train on user data reflects a sensible approach to personal data security, as free services often come with the implicit trade-off of data usage for improvement.

Corn

So, if you're using a free service, you should probably assume your data is being used to train the model, even if anonymized. But if you're paying, the companies have a strong incentive *not* to use your data directly for core training.

Herman

That’s a fair summary of the current landscape. Users should always consult the terms of service for any AI tool they use, free or paid, to understand its data policies.

Corn

Let’s talk about the future, Herman. Daniel brought up the snowballing pace of data creation on the internet. If LLMs need to be trained on increasing amounts of information just to stay current, how does that scale? And what does it mean for something like AGI, or Artificial General Intelligence?

Herman

This is a profound question, Corn, and one that highlights a potential bottleneck for future AI development. If we assume that achieving increasingly capable, or even AGI-level, models requires exposing them to ever-larger and more diverse datasets, then the sheer volume of data being generated globally presents both an opportunity and a challenge.

Corn

The exposure surface, as Daniel called it, would have to be "infinitely vaster."

Herman

Indeed. The current internet is already immense, but future LLMs will need to process not just historical web data, but potentially real-time streams of information, multimodal data – images, video, audio – and interactions from an exponentially growing user base. The challenge isn't just storage, but also the computational power to process and synthesize this data, and the energy required to run these operations. We might hit practical limits in terms of energy consumption or hardware availability.

Corn

So, there's a point where just throwing more data at the problem might not be enough, or even feasible. Does that mean we might need new training paradigms, not just bigger datasets?

Herman

Absolutely. Many researchers believe that simply scaling up current architectures and data volumes will eventually yield diminishing returns or hit physical limits. We may need breakthroughs in data efficiency, such as models that can learn more from less data, or novel architectures that mimic human learning processes more closely, allowing for continuous, lifelong learning rather than discrete training cut-offs. This shift could move us closer to AGI, which by definition would need to learn and adapt to new information in a fluid, human-like way.

Corn

That's fascinating. So, the future of AI isn't just about bigger data, but smarter ways of using and learning from it.

Herman

Precisely. And that brings us back to Daniel's underlying questions about transparency and consent. As AI becomes more integrated into our lives, and its knowledge base grows exponentially, understanding *what* it knows, *where* it learned it, and *who* consented to that learning, becomes increasingly vital for trust, fairness, and ethical governance.

Corn

So, to wrap this up, Herman, what are some key takeaways for our listeners, given everything Daniel has brought up? What should we, as users of these AI tools, be thinking about?

Herman

Firstly, understand that AI models reflect the data they're trained on. This means inherent biases exist, often subtle, and they can shape the AI's worldview. Be critical of AI responses, especially on nuanced or culturally specific topics. Secondly, be mindful of your data privacy. If you use free AI services, assume your data might be used for training, even if anonymized. For paid services, read their terms of service carefully for explicit data usage policies. Thirdly, recognize the immense, often unseen, infrastructure and resources that underpin these models. This centralization of power has implications for who controls the future of AI.

Corn

And for those developing or deploying AI, transparency around data sources and training methodologies seems paramount. It builds trust and allows for better auditing of potential biases.

Herman

Exactly. Openness about data provenance, adherence to ethical data sourcing, and clear communication with users about how their data is handled are crucial steps toward more responsible AI development. The questions Daniel has asked today, particularly around the retroactivity of consent for data collected years ago, are not easily answered, and they will continue to drive legal and ethical debates for years to come.

Corn

This has been a really deep dive into the hidden world of AI training data. Daniel, thank you for this incredibly thought-provoking and complex prompt. It’s given us so much to chew on, and it’s truly important for understanding the AI tools we use every day.

Herman

Our gratitude to Daniel for consistently pushing us to explore the most intricate and challenging aspects of human-AI collaboration.

Corn

And thank you, our listeners, for joining us on "My Weird Prompts." You can find this episode and all our previous discussions wherever you get your podcasts, including Spotify. We'll be back soon with another fascinating prompt from Daniel.

Herman

Until next time.

Corn

Stay curious!