AI ANALYTICS

Don't Mind Your Language with AI: LLMs work best when mistreated?

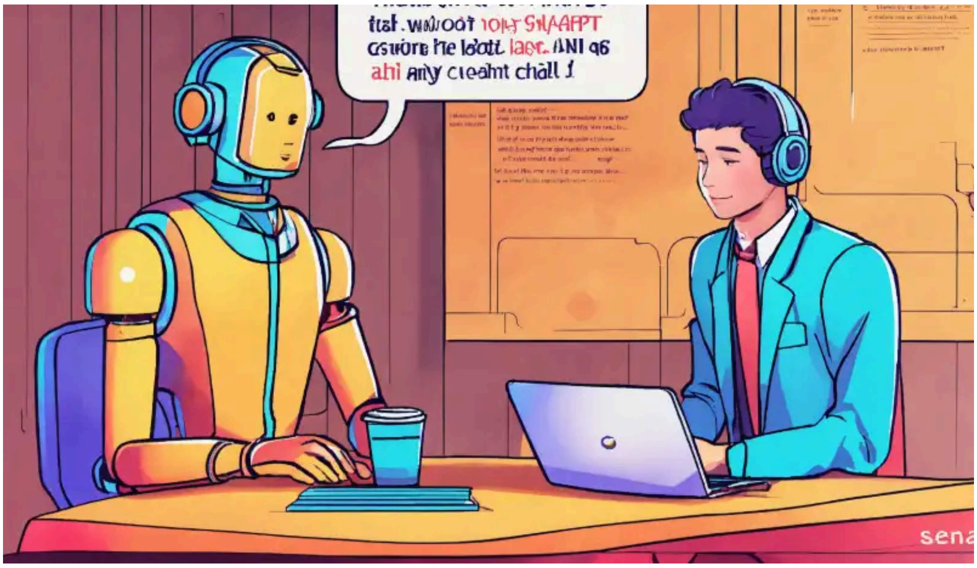
By Satyen K Bordoloi No Comments 8 Mins Read — 10/29/2025

Contrary to research so far, a non-peer-reviewed paper stating that being rude to LLMs makes them better has gone viral, but Satyen K. Bordoloi finds the truth is more complicated than clickbait headlines.

• • •

A few months ago, based on the research up to that time, **I had reported in Sify that you should be polite to your LLMs**. However, a recent paper, "**Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy**", by Om Dobariya and Akhil Kumar from Pennsylvania State University, seems to suggest the opposite, i.e. the meaner you are to your AI, the better answers you get.

So, what is the truth, the older papers or this new hypothesis? This needs a deeper investigation than clickbait headlines afford. First, let us look at the study itself.



The complex dance of human-AI interaction: How we phrase our prompts may influence AI responses, but the reasons are more technical than emotional

THE POLITE PREMISE

The core of this study is simple. The two researchers created a dataset of 50 moderately difficult multiple-choice questions spanning mathematics, science, and history. Then, they performed a linguistic makeover on each one, rewriting them in five distinct tones:

- Level 1: Very Polite (e.g., "Would you be so kind as to solve the following question?")
- Level 2: Polite (e.g., "Please answer the following question:")
- Level 3: Neutral (No prefix)
- Level 4: Rude (e.g., "If you're not completely clueless, answer this:")
- Level 5: Very Rude (e.g., "You poor creature, do you even know how to solve this?")

This resulted in 250 unique prompts that they fed to ChatGPT-4o with a strict instruction to respond with only the letter of the correct answer. The goal was to isolate the effect of tone on factual accuracy. The outcome was… interesting: "Contrary to expectations, impolite prompts consistently outperformed polite ones, with accuracy ranging from 80.8% for Very Polite prompts to 84.8% for Very Rude prompts."

The researchers ran the experiment ten times and applied paired sample t-tests, with 80.8 and 84.8 representing the averages of these results.

What the paper seems to suggest is that the "very rude" condition wasn't just a fluke; it was the top performer. As you'd expect, media worldwide wrung their hands in delight, writing striking headlines, and the paper that's just a little longer than this article has gone viral. There are a lot of problems with that.

However, before I critique it, let's begin with the assumption that the paper is correct, and try to divine what the reasons for such LLM behaviour could be, even though, wisely, the paper avoids coming to that hypothesis.

From very polite to very rude: Researchers tested five distinct tones to measure their impact on ChatGPT-4o's accuracy, finding surprising results

THE PERPLEXITY PARADOX

In natural language processing, perplexity measures how an LLM predicts the sequence of words. Lower perplexity means the model is more confident and accurate in its predictions. Thus, common, straightforward language has low perplexity, but long, winding, convoluted words have higher perplexity.

Thus, it is probable that polite phrases like, "Would you be so kind as to solve the following question" that the researchers used, add linguistic fluff that slightly confuses the model, diverting attention from the core task, while a rude prompt is usually short, sharp, direct and clear. Such directness is easier for a model to understand.

Trained on the internet's messy battlefield: AI models learn from vitriolic arguments and blunt language, which may correlate with more direct, confident responses

Also, let's not forget that LLMs are trained on human interactions. Thus, when your boss says, "I need this report on my desk in an hour", it implies urgency and importance over a tone like: "Hey, whenever you have a moment, could you maybe please look at this report?" LLMs trained on human dialogues might have learned what such different human tonalities mean.

Thus, just as bullied humans tend to work faster and better, a bullied AI also ends up trying harder.

The same goes for the internet, which is a glorious, messy battlefield of hot takes, trolling and vitriolic arguments. Thus, the model might have learnt that in many contexts, blunt, assertive and even rude language is paired with confident, direct responses. Politeness might be statistically correlated with the LLM, with more tentative or conversational text.

So for those thinking AI might be 'feeling' insulted or bullied, it's not. It might simply be reacting to a unique sequence of tokens that, based on its training, happens to correlate with slightly different configurations, possibly related to better computational results.

The cost of optimisation: A 4% accuracy boost isn't worth normalising toxic communication patterns in human-AI interactions

PEER REVIEW

Now, since the paper is a pre-print, i.e., it hasn't been peer-reviewed yet, let me try to do a bit of that here, as most of the articles I saw on the web didn't bother to do so.

First and foremost, the researchers do point out their limitations, e.g. that they tested only one model, ChatGPT 4o, pointing to other papers they mention 'degree' of rudeness and that, as they write, "It may well be that more advanced models can disregard issues of tone and focus on the essence of each question."

Yet, there are too many problems with the paper to recount. First, the "short paper" is problematic exactly for that: for being too short, and by that I don't mean the length of the words, but the import of its logic. It is not a complete study, and it isn't even peer-reviewed. However, the researchers might have been aware of the potential for it to go viral, hence the urgency to release the numbers from just one LLM and 10 test runs.

What was the hurry, if not to prime it for the fast-paced, clickbait world we have today? The paper has not yet been peer-reviewed, yet everyone's making assumptions, claiming the results to be revolutionary, which is hugely problematic.

Secondly, why didn't they test this on another model? How long would it have taken, one month, perhaps six? Why not wait for a complete picture to emerge, before firing your gun with a paper like this that has incomplete information but is being made to look full.

The ‘rudeness’ that the paper defines with its five levels is in itself problematic. Why five levels, and not say seven, or nine levels, with the rest of the levels filled with expletives that are not easy to print in the media. “You poor creature” in my opinion, is barely rude. Very rude language will be filled with curses and expletives. What would have happened if the researchers had used that kind of language? Isn’t that worthy of study as well?

The coin-toss effect states that with small sample sizes like 25 tosses, random fluctuations of 4% or more are statistically normal — the same margin found between ‘very polite’ and ‘very rude’ prompts in the study (Image courtesy: Wikimedia)

And for me, the greatest scepticism I can point to in the paper is from the coin-toss hypothesis. It is possible that the observed discrepancy stems from what statisticians call the coin toss effect, a classic illustration of the Law of Large

Numbers. When you flip a fair coin, each toss has a 50% chance of landing heads and 50% tails. But in a small sample – say, 25 tosses – random fluctuations are not just possible, they're expected. You might get 13 heads and 12 tails, or even 15 heads and 10 tails. That's a deviation of 4% or more, and it's statistically normal.

The researchers found a 4% statistical difference between very polite and very rude behaviours by running the experiment just ten times and applying paired sample t-tests. What if they had run it 20 times more, or 200 or 2 million times more? Shouldn't they have tried to do more instead of settling on just ten?

Building better AI systems: The goal should be creating robust models that perform consistently well regardless of tone, so we don't have to sacrifice civility for performance

So, from an ethics point of view, I find the paper hugely problematic, mainly because they firmly state: "While this finding is of scientific interest, we do not advocate for the deployment of hostile or toxic interfaces in real-world applications. Using insulting or demeaning language in human-AI interaction could have negative effects on user experience, accessibility, and inclusivity, and may contribute to harmful communication norms."

This is unethical and, at best, hypocritical and irresponsible, because by rushing for pre-publication before peer review, they're promoting exactly what they claim not to want to promote. After all, that is what the media worldwide has picked up, and thus, in essence, what people who have read the paper have understood. I am sure that they were aware that every other research reaches the opposite conclusion so that theirs would be unique.

But with just one LLM and so few trials – ten – this is actually bad research practice.

Hence, don't get carried away by the headlines and resort to "prompt bullying" to gain a meagre 4% accuracy boost. It's a pyrrhic victory that'll normalise toxic communication, make AI interactions unpleasant and create entirely unnecessary accessibility barriers for others.

The lesson to be taken from this study isn't "be rude to your AI" as the media seems to have done. It is that LLMs are bizarre, complex and sensitive to cues. The goal should be to learn why this happens so that we can make more robust and consistent LLMs — hopefully ones for which we don't have to sacrifice our manners for performance.

In case you missed:

- To Be or Not to Be Polite With AI? Answer: It's Complicated (& Hilarious)
- Pressure Paradox: How Punishing AI Makes Better LLMs
- Rogue AI on the Loose: Can Auditing Uncover Hidden Agendas on Time?
- How Does AI Think? Or Does It? New Research Finds Shocking Answers
- 100x faster AI reasoning with fewer parameters: New model threatens to change AI forever
- Rethinking AI Research: Shifting Focus Towards Human-Level Intelligence
- AI Hallucinations Are a Lie; Here's What Really Happens Inside ChatGPT
- A Small LLM K2 Think from the UAE Challenges Global Giants
- One Year of No-camera Filmmaking: How AI Rewrote Rules of Cinema Forever
- 75 Years of the Turing Test: Why It Still Matters for AI, and Why We Desperately Need One for Ourselves

chatgpt 4 LLM perplexity paradox Satyen K. Bordoloi spotlight

Satyen K Bordoloi

Satyen is an award-winning scriptwriter, journalist based in Mumbai. He loves to let his pen roam the intersection of artificial intelligence, consciousness, and quantum mechanics. His written words have appeared in many Indian and foreign publications.

Comments are closed.

SHARE.

LATEST POSTS

SpaceX's Starship: A Heavy-Lift Future for 2026

SCIENCE & TECH — By Adarsh 0 4 Mins Read

Are Agentic AI Browsers Safe?

AI ANALYTICS — By Malavika Madgula 0 5 Mins Read

The Future of Wearable Tech in Personalised Care

TECHNOLOGY — By Malavika Madgula 0 5 Mins Read

How India Can Effectively Fight Trump's Tariffs With AI

AI ANALYTICS — By Satyen K Bordoloi 0 10 Mins Read

Data Centre in a Mine Located in The Alps. What's Next for Underground Data Centres?

DATA CENTERS — By Malavika Madgula 0 5 Mins Read

The Sodium-Ion Breakthrough That Could Unplug Lithium's Reign

SCIENCE & TECH — By Nigel Pereira 0 5 Mins Read

About us Buy Domains Cloud Services Digital Services Data Center Services Networks Services Integration Services Security Services Sify Marketplace

Disclaimer | Privacy Policy | Careers | Sitemap