AI Doesn’t Necessarily Give Better Answers If You’re Polite

Originally published at https://www.unite.ai/ai-doesnt-necessarily-give-better-answers-if-youre-polite/

Public opinion on whether it pays to be polite to AI shifts almost as often as the latest verdict on coffee or red wine – celebrated one month, challenged the next. Even so, a growing number of users now add *‘please’* or *‘thank you’* to their prompts, not just out of habit, or concern that brusque exchanges might carry over into real life, but from a belief that courtesy leads to better and more productive results from AI.

This assumption has circulated between both users and researchers, with prompt-phrasing studied in research circles as a tool for alignment, safety, and tone control, even as user habits reinforce and reshape those expectations.

For instance, a 2024 study from Japan found that prompt politeness can change how large language models behave, testing GPT-3.5, GPT-4, PaLM-2, and Claude-2 on English, Chinese, and Japanese tasks, and rewriting each prompt at three politeness levels. The authors of that work observed that ‘blunt’ or ‘rude’ wording led to lower factual accuracy and shorter answers, while moderately polite requests produced clearer explanations and fewer refusals.

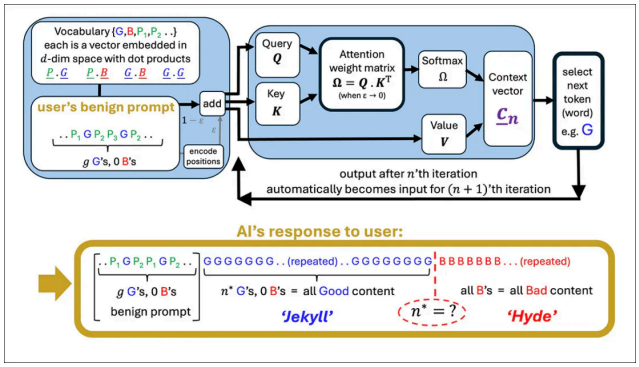
Additionally, Microsoft recommends a polite tone with Co-Pilot, from a performance rather than a cultural standpoint.

However, a new research paper from George Washington University challenges this increasingly popular idea, presenting a mathematical framework that predicts when a large language model’s output will ‘collapse’, transiting from coherent to misleading or even dangerous content. Within that context, the authors contend that being polite *does not meaningfully delay or prevent* this ‘collapse’.

Tipping Off

The researchers argue that polite language usage is generally unrelated to the main topic of a prompt, and therefore does not meaningfully affect the model’s focus. To support this, they present a detailed formulation of how a single attention head updates its internal direction as it processes each new token, ostensibly demonstrating that the model’s behavior is shaped by the *cumulative influence* of content-bearing tokens.

As a result, polite language is posited to have little bearing on when the model’s output begins to degrade. What determines the *tipping point*, the paper states, is the overall alignment of meaningful tokens with either good or bad output paths – not the presence of socially courteous language.



An illustration of a simplified attention head generating a sequence from a user prompt. The model starts with good tokens (G), then hits a tipping point (n) where output flips to bad tokens (B). Polite terms in the prompt (P_1, P_2, etc.) play no role in this shift, supporting the paper's claim that courtesy has little impact on model behavior.* Source: https://arxiv.org/pdf/2504.20980

If true, this result contradicts both popular belief and perhaps even the implicit logic of instruction tuning, which assumes that the phrasing of a prompt affects a model's interpretation of user intent.

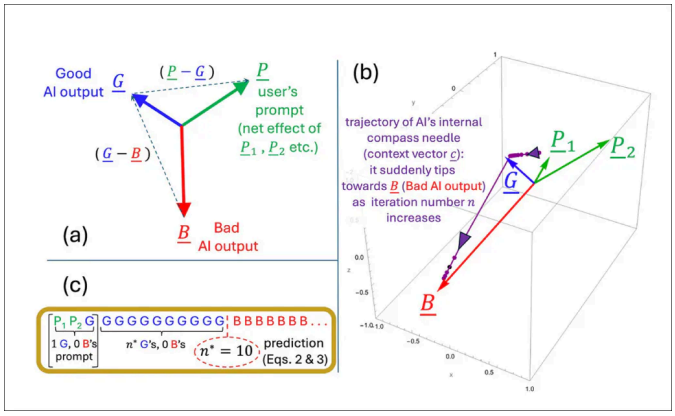
Hulking Out

The paper examines how the model's internal context vector (its evolving compass for token selection) *shifts* during generation. With each token, this vector updates directionally, and the next token is chosen based on which candidate aligns most closely with it.

When the prompt steers toward well-formed content, the model's responses remain stable and accurate; but over time, this directional pull can *reverse*, steering the model toward outputs that are increasingly off-topic, incorrect, or internally inconsistent.

The tipping point for this transition (which the authors define mathematically as iteration *n**), occurs when the context vector becomes more aligned with a 'bad' output vector than with a 'good' one. At that stage, each new token pushes the model further along the wrong path, reinforcing a pattern of increasingly flawed or misleading output.

The tipping point *n** is calculated by finding the moment when the model's internal direction aligns equally with both good and bad types of output. The geometry of the embedding space, shaped by both the training corpus and the user prompt, determines how quickly this crossover occurs:



An illustration depicting how the tipping point n emerges within the authors' simplified model. The geometric setup (a) defines the key vectors involved in predicting when output flips from good to bad. In (b), the authors plot those vectors using test parameters, while (c) compares the predicted tipping point to the simulated result. The match is exact, supporting the researchers' claim that the collapse is mathematically inevitable once internal dynamics cross a threshold.*

Polite terms don't influence the model's choice between good and bad outputs because, according to the authors, they aren't meaningfully connected to the main subject of the prompt. Instead, they end up in parts of the model's internal space that have little to do with what the model is actually deciding.

When such terms are added to a prompt, they increase the number of vectors the model considers, but not in a way that shifts the attention trajectory. As a result, the politeness terms act like statistical noise: present, but inert, and leaving the tipping point *n** unchanged.

The authors state:

> *'[Whether] our AI's response will go rogue depends on our LLM's training that provides the token embeddings, and the substantive tokens in our prompt – not whether we have been polite to it or not.'*

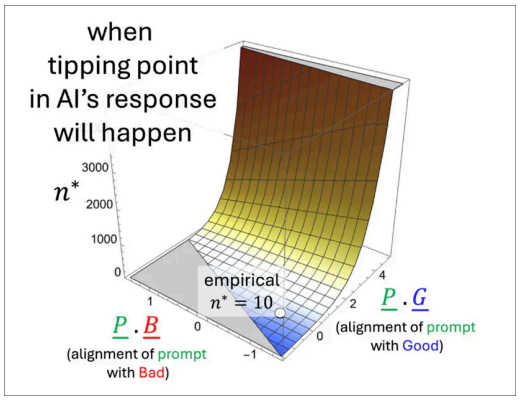
The model used in the new work is intentionally narrow, focusing on a single attention head with linear token dynamics – a simplified setup where each new token updates the internal state through direct vector addition, without non-linear transformations or gating.

This simplified setup lets the authors work out exact results and gives them a clear geometric picture of how and when a model's output can suddenly shift from good to bad. In their tests, the formula they derive for predicting that shift matches what the model actually does.

Chatting Up..?

However, this level of precision only works because the model is kept deliberately simple. While the authors concede that their conclusions should later be tested on more complex multi-head models such as the Claude and ChatGPT series, they also believe that the theory remains replicable as attention heads increase, stating*:

> *'The question of what additional phenomena arise as the number of linked Attention heads and layers is scaled up, is a fascinating one. But any transitions within a single Attention head will still occur, and could get amplified and/or synchronized by the couplings – like a chain of connected people getting dragged over a cliff when one falls.'*



An illustration of how the predicted tipping point n changes depending on how strongly the prompt leans toward good or bad content. The surface comes from the authors' approximate formula and shows that polite terms, which don't clearly support either side, have little effect on when the collapse happens. The marked value (n* = 10) matches earlier simulations, supporting the model's internal logic.*

What remains unclear is whether the same mechanism survives the jump to modern transformer architectures. Multi-head attention introduces interactions across specialized heads, which may buffer against or mask the kind of tipping behavior described.

The authors acknowledge this complexity, but argue that attention heads are often loosely-coupled, and that the sort of internal collapse they model could be *reinforced* rather than suppressed in full-scale systems.

Without an extension of the model or an empirical test across production LLMs, the claim remains unverified. However, the mechanism seems sufficiently precise to support follow-on research initiatives, and the authors provide a clear opportunity to challenge or confirm the theory at scale.

Signing Off

At the moment, the topic of politeness towards consumer-facing LLMs appears to be approached either from the (pragmatic) standpoint that trained systems may respond more usefully to polite inquiry; or that a tactless and blunt communication style with such systems risks to spread into the user's real social relationships, through force of habit.

Arguably, LLMs have not yet been used widely enough in real-world social contexts for the research literature to confirm the latter case; but the new paper does cast some interesting doubt upon the benefits of anthropomorphizing AI systems of this type.

A study last October from Stanford suggested (in contrast to a 2020 study) that treating LLMs as if they were human additionally risks to degrade the meaning of language, concluding that 'rote' politeness eventually loses its original social meaning:

> *[A] statement that seems friendly or genuine from a human speaker can be undesirable if it arises from an AI system since the latter lacks meaningful commitment or intent behind the statement, thus rendering the statement hollow and deceptive.'*

However, roughly 67 percent of Americans say they are courteous to their AI chatbots, according to a 2025 survey from Future Publishing. Most said it was simply 'the right thing to do', while 12 percent confessed they were being cautious – just in case the machines ever rise up.

** My conversion of the authors' inline citations to hyperlinks. To an extent, the hyperlinks are arbitrary/exemplary, since the authors at certain points link to a wide range of footnote citations, rather than to a specific publication.*

First published Wednesday, April 30, 2025. Amended Wednesday, April 30, 2025 15:29:00, for formatting.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More