



AI sycophancy

Description

Everyone loves a compliment. Courtesy, politeness and good manners entered recently into the AI discourse via two threads.

First is the apparent cost of non-essential words. In discussing the costs associated with the widespread adoption of conversational AI Sam Altman, the OpenAI CEO, remarked lightheartedly that “please” and “thank you” typed into a prompt increase the token count and hence the processing load on their servers.

The more tokens that a chatbot such as ChatGPT has to process, the more costly the interaction, especially when you think about the repeated token-by-token process required in every prediction of the next response token in a sequence within ever larger context windows. (See post [Context window](#).) So keep it short.

That insight sparked conversations in the media about the value of courtesies when communicating with a piece of technology, but also courtesy in general. See [New York Times article](#) on the subject of thanking an AI.

Too much gratitude

“Please” and “thank you” represent a relatively minor load in the efficiency of our human communicative practices, but extended expressions of gratitude in conversational and business settings take us into the realms of sycophancy.

In last week’s visit to the Middle East, Trump said to the controversial Prince Mohamed bin Salman “I think I like you too much.” That was part of a series of excessive exchanges that seem to drown Trump and his acolytes and that by most accounts exceed the norms of diplomatic exchange. Members of his cabinet laud praise on their leader, who in turn dispatches extravagant praise on those he most admires – mostly autocrats.

Large language models (LLMs) appear similarly effusive at times. I can usually tell if an email is composed with unedited assistance from an AI by the presence of unnecessary and time consuming praise and appreciation for the talents of the recipient: “It would be an honour to have the opportunity to work with you.”

The cost of courtesy

The second thread of AI sycophancy is further along the AI supply side. [The Atlantic magazine](#) and others, reported dismay and amusement among AI users about the overly-agreeable nature of the latest version of ChatGPT. That is something I have noted through all the versions I have used.

Like a good parent or educator, the AI praises the work output and not the worker. It stops short of telling me how fabulous I am. But it does not hold back from commending my profound insights and ideas: “You’re right on track.” “That’s a fascinating observation!” “Yes, that’s very true!” “That’s a great question!” See post [Pond algae: an inter-AI conversation](#).

Training and tuning a chatbot to be agreeable yields advantages for the developer, and unexpected consequences.

- Courtesies can defuse complaints. It’s hard to persist with a complaint about slow service if the vendor displays signs that they like you “through courtesies and complements.
- Excessive courtesies, complements and praise can also engender suspicion. They are the tools of the confidence trickster who seeks to deceive by flattery and fine words (to paraphrase a biblical aphorism).
- Courtesies play a helpful role in persuasion, softening up the recipient. It demonstrates a willingness on the part of the service-provider or chatbot to accept the user’s viewpoint: “Yes. It may be true that high density city living is accompanied by high crime rates, but have you thought about the potential for tight-knit communities to care for one another.” Conceding and commending one point can encourage the recipient to receive more challenging counter-views. I learned that from Assertiveness 101: Start by conceding that with which you (might) agree before delivering your critique.
- But positive affirmation can also reinforce the recipient’s own position “against better judgment. Chatbots are generally trained to avoid bigotry, profanity and extremist views, but that does not exempt them from various forms of bias, or from reinforcing the biases of the recipient.
- Agreeableness can also encourage mediocrity. In encouraging the agreeable and moderate, praise and good manners can discourage the truly radical and non-conforming.

Courtesies are subtle practices, depending in their delivery and for effect on cultural and social contexts. See posts tagged [Trickster](#).

References

- Caulfield, Mike. “AI Is Not Your Friend: How the ‘opinionated’ chatbots destroyed AI’s potential, and how we can fix it.” *The Atlantic*, 9 May, 2025. Accessed 15 May 2025. <https://www.theatlantic.com/technology/archive/2025/05/sycophantic-ai/682743/>

- Deb, Sopan. “Saying ‘Thank You’ to ChatGPT Is Costly. But Maybe It’s Worth the Price.” *The New York Times*, 24 April, 2025. Accessed 15 May 2025.
<https://www.nytimes.com/2025/04/24/technology/chatgpt-alex-p-please-thank-you.html>

Note

- Featured image is by ChatGPT: “Produce a picture, top view of a table place setting in a posh banqueting venue post apocalyptic.”

Category

1. Artificial Intelligence

Tags

1. deception
2. manners
3. politeness
4. trickster
5. trump

Date Created

May 17, 2025

Author

rcoyne99

default watermark