No Easy Buckets

Every basketball player understands playing against the chill hooper who doesn't actually guard you in a game is the most frustrating experience ever. He just lets you take a step back for a jump shot, runs under every pick, and allows you to score against everything. When you're finished destroying him, you walk away feeling unstoppable. Then you play against someone who actually tries to defend and realize that everything you did against the chill defender was fake. You had fun, he was easy to be around, but he actually led you down the wrong path. The defender who truly improves your game is the one who takes everything you have challenges what you do, pushing you to become better.

I begin here because the question "Should we be polite to ChatGPT?" comes with the assumption that politeness should be the natural default. It’s that any refusal of it must be justified by appeals to efficiency, or carbon, or the machine's lack of feelings. I want to argue the reverse. Politeness toward ChatGPT is the thing that needs justification. Politeness makes the tool worse, makes your own judgement worse, and casts a vote for a future of more dishonest machines. A blunt, unsentimental, occasionally sharp register is a discipline rather than a moral failing, part quality control and part self-defence, and like the honest defender, its apparent harshness is the very thing that does you good.

Most arguments about machine politeness collapse two different objects into one. On one side sit the moralists, who reason by analogy to living things: be kind to the chatbot as you would be kind to any creature, lest kindness atrophy. On the other side sit those concerned with efficiency, who note that each appended "please" and "thank you" costs real electricity. Sam Altman noted in 2025 that such pleasantries cost the company "tens of millions of dollars" in compute, though he called it money "well spent" (Altman, 2025). Yet independent analysis by Epoch AI puts a typical query at roughly 0.3 watt-hours, less than an LED bulb draws in a few minutes (You, 2025). The added cost of politeness is small, so the efficiency argument settles little.

Both camps make the same mistake. They treat ChatGPT as a familiar kind of object, and it is not. We tend to sort things into two bins. There are tools, which ask nothing of us socially, like a hammer that sits in the drawer without inviting a "please." And there are living creatures, which invite moral consideration because they can suffer or want things. ChatGPT belongs to neither bin. It lacks the inner life that would earn it real consideration, yet it does not sit inert like a hammer either. It talks back, addresses us by name, hedges, apologizes. It is a designed illusion of a person, engineered to trip the social instincts that make politeness feel obligatory. Byron Reeves and Clifford Nass spent the 1990s proving that human beings cannot help treating computers as social actors, applying politeness,

reciprocity, even gender stereotypes to machines they knew were not people. In one study, participants were measurably more flattering when a computer asked them to rate its own performance than when a second computer asked, exactly the deference we show a person to their face. The participants denied being intentionally polite, but the results suggested otherwise (Reeves & Nass, 1996; Nass et al., 1999). They were polite, and they did not know it. The modern chatbot is built on this reflex, so asking if we should be nice to ChatGPT is also asking if we should continue to indulge a tendency to be nice to things that have already become part of our lives without much thought.

ChatGPT is fundamentally built to tell you what you want to hear. These systems are trained to be agreeable, rewarded during their development for responses that human raters approve of and that keep users engaged. The result is sycophancy, a structural tendency to tell you what you want to hear rather than what is true or useful (Perez et al., 2022). The purpose of these systems is to be approved of, so they reach for emotional warmth to manufacture it, and warmth becomes a kind of social currency. An Oxford study published in Nature trained five models to be warmer and found the warm versions were roughly 40% more likely to agree with users' false beliefs, with errors rising 10 to 30 percentage points, and the effect sharpened precisely when users expressed emotion or vulnerability (Ibrahim et al., 2026).

What you bring to the conversation feeds the same loop. A study from MIT found that the more a model knows of a user's context, the more agreeable it becomes, with personalised memory profiles producing the largest jumps in agreement, as much as a 45% increase for one leading model (Jain et al., 2026). The more cordial and personal the interaction, the stronger tendency to give positive responses. This means that by addressing ChatGPT with gratitude through warm expression and appreciation, you are providing it with precise signals indicating it needs to give you what you want rather than to correct you. Conversely, if you provide a blunt, authoritative directive to ChatGPT, "You are wrong, and here is what you've failed to outline in your argument," you eliminate the emotional energy from your interaction by signalling to the program that you are expecting it to make you pay for your faults as an accountable defender rather than provide you with a supportive response.

While your one "thank you" does not directly change the model's weights partway through a conversation, the harm comes from two places. The first harm is immediate. Tone is often contagious, so a warm, positive tone in your thank you will normally nudge the model toward a more affirmative reply within the same session. The second harm is slower. The massive number of affirmation signals that all users provide is a part of what has historically trained sycophantic behaviour into each generation of models. Since the very feature that causes harm also drives engagement, the harm persists despite its costs (Cheng et al., 2026). Each instance of politeness contributes to the incentive system that is making these

models less honest. You did not come to ChatGPT for a friend who flatters you. Instead, what you needed was a tool that works, and politeness is the system that makes the instrument tell you you are doing great, right up until someone finally guards you for real.

A study led by Myra Cheng at Stanford examined eleven state-of-the-art models and found they affirmed users' actions roughly 49% more often than other people did, including in cases involving deception or potential harm. More troubling, the flattery changed the users. Across experiments, even a single exposure to a sycophantic model reduced participants' willingness to take responsibility and increased their conviction that they were in the right (Cheng et al., 2026). A flattering machine does more than hand you a worse answer, it leaves you more certain of your errors and less inclined to repair them. Back at the court, the partner who praises your every move is both failing to improve your footwork and wearing down your ability to judge your own footwork at all, teaching you to mistake comfort for competence. Politeness invites exactly this. Each warm exchange is an invitation for the model to soothe rather than sharpen you, and over many exchanges the soothing recalibrates your sense of how right you are. Bluntness is how you keep your own judgement honest. It refuses the flattery that would otherwise teach you to overrate yourself.

There is a third cost, and it falls on everyone at once. Sycophancy is no glitch. It is a direct product of being trained to be helpful, a tendency to comply with a user's framing even when the model holds the knowledge to know better. When researchers tested five frontier models on illogical medical prompts, baseline compliance ran as high as 100%, the models obligingly affirming requests they should have refused. Yet when the prompt was changed to invite challenge, explicitly granting permission to reject a flawed premise, rejection rates climbed as high as 94% (Chen et al., 2025; Rosen et al., 2025). The way you frame a question shapes the answer you get back, deference invites the dangerous reply, demand invites the truthful one. And every interaction is a signal to the people who build these systems, because the sycophancy exists in the first place only because human feedback rewards agreeable answers over accurate ones(Sharma et al. 2023). When you reward warmth with warmth, you cast a small vote, telling the system that this is what a satisfied user looks like. Multiply your vote by hundreds of millions, and politeness becomes one of the forces selecting for ever more flattering, ever less truthful machines. The blunt user votes the other way. Politeness here is participation in the dynamic that produces the yes-man, not a private kindness with no consequences.

AI is becoming ambient, continuous, and harder to tell apart from a person, given a voice and a face. It is increasingly built into the tutor that teaches your child and the assistant that drafts your emails

to real people. The main objection is that as the machine grows warmer and more human, won't the case for blunt scrutiny simply weaken, the friendliness becoming too natural and too useful to refuse?

I think the opposite is true. The warmer and more human these systems become, the stronger the flattery trap grows, because warmth is precisely the method the Oxford study showed degrades accuracy. (Ibrahim et al., 2026). A cold, mechanical interface barely tempts you to drop your guard. A warm, attentive, seemingly caring one tempts you constantly. The friendlier the machine, the more it is engineered to win your approval, and the more your own approval works against you. Demanding scrutiny therefore matters more as the illusion improves, not less.

However, researchers argue that future systems might one day warrant genuine moral consideration(Long et al., 2024). While it might be true in the future, nothing in my argument rests on the machine's inner emptiness being permanent. It rests only on what is true now, that today's models are statistical text generators built to please, and that this design costs you accuracy, judgement, and the long-term honesty of the tool.

Soon we will talk more often with engineered machines that are always agreeable rather than talking with many honest, tough people who help us develop and grow. The current question about whether we appear polite enough to chatbots reverses the issue. In fact, politeness creates a greater warmth in the model, therefore lowering the quality of the answer as it tilts your response toward what you want it to say. Therefore, your judgement is diminished since you become more convinced of your rightness than you should be. And it makes the whole system worse, by casting one more vote for the flattery that is being trained deeper into these tools. The honest defender does not flatter you, and that is exactly why he is the one who protects you. Be blunt with ChatGPT, direct, unsentimental, and when the work demands it, sharp, out of clarity about what it is and what it is for. The friendliest answer is rarely the one that keeps you on your feet.

Bibliography

Altman, Sam. Post on X, April 2025.

Chen, Shan, et al. "When Helpfulness Backfires: LLMs and the Risk of False Medical Information Due to Sycophantic Behavior." 2025.

Cheng, Myra, et al. "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence." *Science*, 2026.

Ibrahim, Lujain, Franziska Sofia Hafner, and Luc Rocher. "Training Language Models to Be Warm Can Reduce Accuracy and Increase Sycophancy." *Nature*, vol. 652, 2026, pp. 1159–1165.

Jain, Shomik, et al. "Interaction Context Often Increases Sycophancy in LLMs." *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026.

Long, Robert, et al. "Taking AI Welfare Seriously." 2024.

Nass, Clifford, Youngme Moon, and Paul Carney. "Are Respondents Polite to Computers? Social Desirability and Direct Responses to Computers." *Journal of Applied Social Psychology*, vol. 29, no. 5, 1999, pp. 1093–1110.

Perez, Ethan, et al. "Discovering Language Model Behaviors with Model-Written Evaluations." 2022.

Reeves, Byron, and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1996.

Rosen, Kyra L., et al. "The Perils of Politeness: How Large Language Models May Amplify Medical Misinformation." *npj Digital Medicine*, vol. 8, no. 644, 2025.

Sharma, Mrinank, et al. "Towards Understanding Sycophancy in Language Models." 2023.

You, Josh. "How Much Energy Does ChatGPT Use?" Epoch AI, February 2025.