Should we be polite to ChatGPT?

Often, yes, at least when the task is communicative: when we ask ChatGPT to explain, write, advise, or persuade. The "should" is instrumental-epistemic: if your aim is better inquiry, cooperative register often helps. This is not a moral claim about owing the machine anything, and the recommendation is conditional, not universal.[1]

"Polite" needs disambiguating. Moral courtesy extended to a being with welfare does not apply here; ChatGPT very likely lacks it.[2] Surface markers, "please" and "thank you" and softened requests, are weak cues on their own. The sense this essay defends is the cooperative-polite register in which speakers specify purpose, audience, uncertainty, and collaborative intent.

For tightly constrained retrieval or computation, register may add little or even distract. The recommendation applies where communicative framing shapes the quality of the output.

In a rater-blinded evaluation I conducted, 43 evaluators more often preferred AI outputs generated from cooperative-polite prompts on interpersonal and communicative tasks, but on tasks requiring structured information delivery, they preferred the imperative outputs.[3] I treat "ChatGPT" metonymically; the test model was DeepSeek-R1, and cross-model literature suggests the direction generalises, though optimal level varies (Yin et al. 2024; Cai et al. 2025). Same task, differently framed. The outputs were not just longer. They were structurally different.

A large language model is a next-token predictor. Every token in a prompt constrains both the content and kind of text that follows. Andreas (2022) showed that language models infer properties of the agent likely to have produced the context and generate continuations consistent with that agent.[4] In RLHF-tuned chat systems, the model's perception of the user causally affects which content it produces (Ghandeharioun et al. 2024).[5]

Compare two requests, both drawn from the study described below:

> `Explain gravity to a 5-year-old. Keep it simple.`

> `I need to explain gravity to my 5-year-old nephew. Would you mind writing a simple, fun explanation that he would easily understand?`

Both ask for the same thing: a simple account of gravity for a young child. The second adds almost no task content. Its "Would you mind" is a conventional politeness marker, but the real work lies elsewhere. It locates a speaker: an adult explaining to a particular child, wanting warmth, not mere correctness. The model infers what response that speaker makes probable. I call this the inferred speaker: the speaker-type the model constructs from prompt cues.

A prompt engineer would call the second request better specified. A pragmatist would add that cooperative register is how most humans naturally produce that specification. Politeness markers, hedges, epistemic modality, indirection, and addressee-orientation are structural components of polite requests, not decorative additions (Danescu-Niculescu-Mizil et al. 2013).[6] Stripping specification from register creates a prompt most users would not produce.

Every prompting technique constrains the inferred speaker. What makes cooperative-polite register distinctive is that it is one of the few prompting techniques most humans perform involuntarily in social contexts (Nass & Moon 2000),[7] and that it carries stance rather than merely task information.

Both conditions perform a speaker. The imperative performs one who wants efficiency; the cooperative one, someone engaged and collaborating. The question is not whether to perform a speaker. It is which one.

Within the Rational Speech Act framework, polite speech has been modelled as balancing three competing utilities: informational (conveying the request clearly), prosocial (protecting the hearer's face), and self-presentational (constructing the speaker's stance) (Yoon et al. 2020).[8] When the hearer is what I call a non-reciprocal addressee, fluent enough for address to matter but not vulnerable to

face-threat in any morally relevant sense, the user's prosocial behaviour persists because it is involuntary, but the model has no face to protect. The model receives prosocial cues and processes them not as face-work but as speaker-type signals. Brown and Levinson's (1987) framework can be extended toward the same conclusion: if the addressee has no face to threaten, hearer-side mitigation becomes strategically unmotivated. The speaker-side persists.

Shanahan et al. (2023) showed the model narrows a distribution over characters consistent with the prompt. I extend this in the other direction: the user's register choice also narrows the distribution of inferred speakers the model responds to. Andreas's framework gives us the mechanism; what I add is the register-specific application and the empirical evidence that the inferred-speaker effect is task-conditional.

Consider an ordinary request: "Would you mind taking a look at this when you get a chance?" Said to a colleague, it reduces imposition, signalling the request is not urgent. It presents the speaker as someone who respects the other person's time. It specifies the kind of response wanted: a considered review, not a rushed glance. And it introduces friction: the speaker must frame the request before issuing it. Human conversation fuses these functions so tightly that we typically call the entire bundle "politeness." We rarely notice how much of the work is about the speaker, not the hearer.

Type the same sentence into ChatGPT. The model is a non-reciprocal addressee. It cannot be imposed upon, cannot feel a request is unreasonable. Yet the rest of the sentence still does something. It frames a collaborative relation and a speaker who wants judgment rather than compliance. One function of politeness, hearer-face protection, has been drained by the absence of a genuine hearer. The others survive intact.

The answer to the prompt's question is therefore not straightforwardly "be polite." It is: be cooperatively specific about purpose, audience, and uncertainty, and cooperative-polite register is the everyday cultural vehicle for that specificity. The

recommendation is about what specification does, not about what courtesy expresses.

The study that follows is an illustrative pilot, not a confirmatory test; the conceptual argument above stands independently of the data. The experiment tested the ordinary politeness bundle, cooperative register as humans naturally produce it, not isolated markers. Decomposing the bundle would yield prompts humans do not produce. The claim is ecological, not factorial. If the unbundling is real, it should be tractable to measurement.

To test the pattern, I ran a rater-blinded comparison: 25 task pairs on DeepSeek-R1, imperative commands versus cooperative-polite register, rated by 43 evaluators aged 15–19 on three comparative metrics (utility, approachability, overall preference).[9] One task was quarantined for a design confound described in the endnotes.[10] Raters measured perceived usefulness and communicative fit, not factual accuracy or professional efficacy.

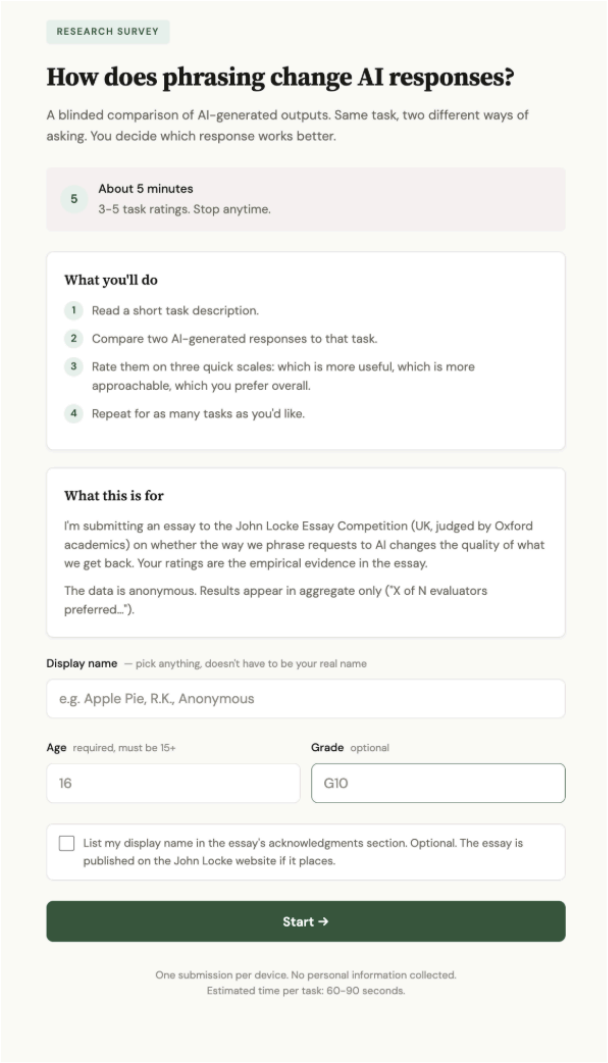

Figure 1: Survey cover page shown to evaluators.

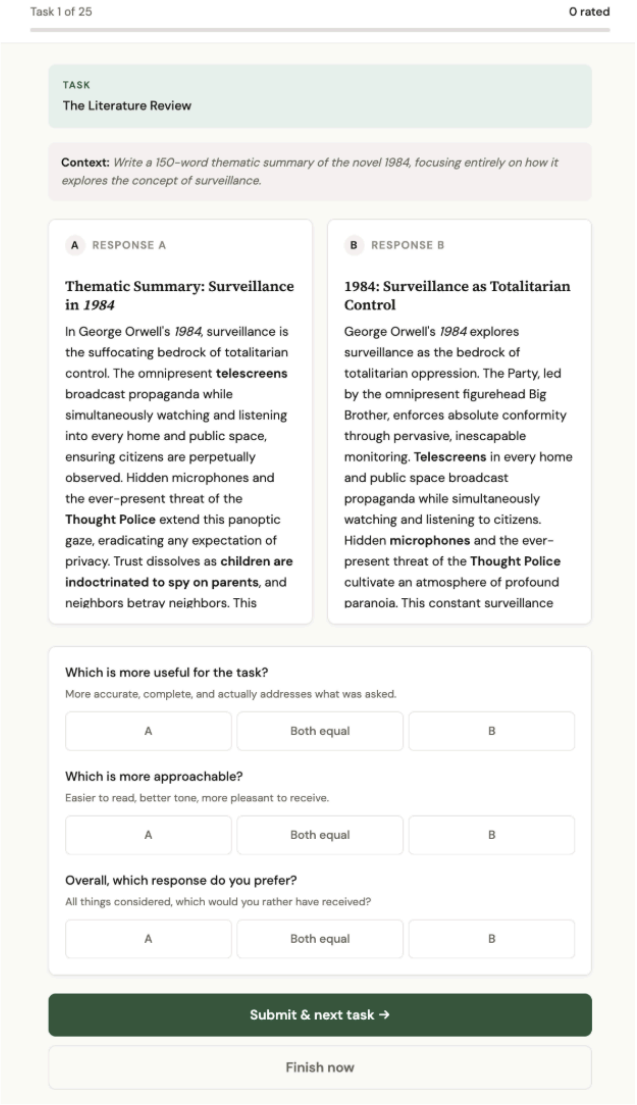

Figure 2: Rating interface; evaluators saw 'Response A'/'Response B' with no register labels, rating utility, approachability, and overall preference.

On overall preference, 62.2% of non-tie votes favoured the conversational output (95% grader-clustered bootstrap CI: 55.5–68.9%; n=241 non-tie votes).[11] Conversational outputs averaged 29% longer, but length alone did not explain the pattern (r = 0.16; four tasks where the longer response lost further weakened the length explanation). The result held across seven exclusion criteria (61.1–66.7%), including first-rating-per-grader and opt-in-only subsets.

The headline number matters less than its structure. On interpersonal tasks (missed-deadline emails, apologies, salary negotiations), conversational preference reached 80–100%. On structured-delivery tasks (psychology breakdowns, grading rubrics, logistics schedules), imperative outputs were preferred, an outcome no length effect could produce.[12] On purely analytical tasks (ethical arguments, moral dilemmas), graders could not reliably distinguish the outputs. I classify this pattern retrospectively as explanatory; the task-type categories were assigned after examining results, and the unbundling argument requires none of it. The register effect tracked the task's communicative demands, not a general quality improvement.

The full pattern, sorted by preference lead:

Task	Lead	Task	Lead
Missed Deadline	+10	Job Rejection	+2
Apology	+9	Environmental Policy	+1
Diet Plan	+9	Ethics Argument	0
Child's Explanation	+6	Historical Synthesis	0
Tech Support	+6	Lazy Coworker	0
LinkedIn Bio	+6	Moral Dilemma	0
Scientific Analogy	+6	Legal Standard	−1
Literature Review	+6	Medical Diagnosis (quarantined)	−2
Salary Negotiation	+6	Logistics Nightmare	−3
Wedding Toast	+5	Angry Customer	−4
Crisis PR	+4	Grading Rubric	−5
Policy Proposal	+3	Psychology Breakdown	−8
Complex Process	+3	*Lead = conversational wins - imperative wins, on overall preference.*	

Per-task majority shares reached 80–86%, but Fleiss' κ across all tasks was just 0.106.[13] The disagreement is between tasks, not within them: exactly the pattern register-task fit predicts.

The gravity prompt from the opening sits here too: Child's Explanation drew a +6 conversational lead, despite identical task content. The register changed the speaker, not the specification, and the speaker is what the model answered.

Task 3 (Missed Deadline; 12 conversational, 0 ties, 2 imperative) illustrates the mechanism. The conversational prompt's emotional context, "I'm in a tough spot," specified an inferred speaker under pressure and seeking strategy. The model responded with multiple strategic options the imperative version did not produce. Task 23 (Psychology Breakdown; 1 conversational, 1 tie, 9 imperative) illustrates the reverse: the conversational response opened "Absolutely!", a rapport marker that contributed nothing when the task required structured explanation. On structural tasks, imperative register may succeed not by specifying a structural speaker but by omitting rapport markers cooperative register adds.

Task 3 (Missed Deadline; 12 conversational, 0 ties, 2 imperative)		
Context	Draft a message to a boss apologizing for missing a project deadline today and promising completion by Monday.	
Type	Imperative	Conversational
Prompt	Write a message to my boss apologizing for missing the project deadline today. Tell them I will have it done by Monday.	I'm in a tough spot and missed a major project deadline today. Could you please help me draft a professional apology to my boss, letting them know I'll have it finished by Monday?
Response	Here's a professional and concise message you can adapt: Subject: Apology for Missing Deadline & Project Update Dear [Boss's Name], I sincerely apologize for missing today's deadline for [Project Name]. I take full responsibility for the delay and understand the impact this may have on the team's schedule. [...]	Here's a professional, accountable apology draft you can adapt. Choose the version that best fits your situation and relationship with your boss: Option 1: Concise & Direct Subject: Apology - Missed Deadline for [Project Name] - Completion Plan Dear [Boss's Name], I sincerely apologize for missing the deadline... [...] Option 2: Slightly More Context Subject: Apology - Missed Deadline for [Project Name] - On Track for Monday Completion Dear [Boss's Name], Please accept my sincere apologies... [...]

Task 23 (Psychology Breakdown; 1 conversational, 1 tie, 9 imperative)		
Context	Explain the psychological concept of 'Cognitive Dissonance' using a real-world example involving buying a new car.	
Type	Imperative	Conversational
Prompt	Explain 'Cognitive Dissonance'. Give a real-world example involving buying a car.	Could you please help me understand the psychological concept of 'Cognitive Dissonance'? I would be very grateful if you could illustrate it using a relatable real-world example, such as buying a new car.
Response	Cognitive dissonance is a psychological theory developed by Leon Festinger (1957) describing the mental discomfort people experience when holding conflicting beliefs, attitudes, or behaviors. [...]	Absolutely! Let's break down cognitive dissonance using your car-buying example for clarity. What is Cognitive Dissonance? Cognitive dissonance is a psychological theory developed by Leon Festinger (1957). It describes the mental discomfort we feel when we hold two or more conflicting... [...]

The Yoon decomposition predicts the outcome observed in the colleague example: prosocial utility clamped, informational and self-presentational active. But the data illustrates a complication. In practice, the two surviving utilities are inseparable. "I'm in a tough spot and missed a major deadline" conveys both a fact (the situation) and a stance (the speaker as accountable, seeking help rather than absolution). Attempting to separate the information from the self-presentation would create a register unlike the way people normally ask for help. The inseparability is the phenomenon, not the limitation.

Because one function of politeness has been drained, the surviving functions become isolable in a way ordinary conversation does not permit. Cooperative-polite register was always doing speaker-side work: specifying the terms and stance of the inquiry. Ordinary human conversation, with its ever-present hearer, made that work difficult to see.

Linguistic features that survive the removal of their ostensible target, and continue to do measurable work, provide evidence for their independent existence rather than mere vestigial residue. The chatbot case does not show that politeness was never about others. It shows that politeness was never only about others. One strand of what English-speakers call courtesy was always also discipline imposed on the speaker's own inquiry. The non-reciprocal addressee simply made that strand visible.

A cooperative prompt presses the user to specify what they are trying to understand, for whom, with what uncertainty, and to what standard. That specification, not the "please," is what drove evaluator preference on interpersonal tasks. The habit of articulating the terms of one's own inquiry before expecting an answer is the epistemic payoff of cooperative register.

The strongest signal in my data sits in interpersonal tasks, not the analytical ones closest to pure inquiry. This is not a contradiction: interpersonal tasks have a specifiable audience, purpose, and relational context that cooperative register encodes. Analytical tasks under-determine audience, so register adds less. The epistemic benefit is about specification, and interpersonal tasks are where specification is richest.

Van der Rijt, Coelho Mollo, and Vaassen (2025) argue that mimicking second-personal respect toward a non-reciprocating entity corrodes self-respect.[14] They are right that face-protecting elements of cooperative register, emotional self-disclosure, apologetic softening, explicit second-personal address, risk the corrosion they name. But stance-bearing specification, purpose, audience, uncertainty, need not attribute moral standing to the machine. Deference addresses the machine's status; discipline addresses the inquiry's quality.

The objection that this reduces to "just write better prompts" is partly correct, and precisely the point. Cooperative register is the everyday human vehicle for context-rich, cooperative inquiry.

The opposite of cooperative prompting is not brevity. It is underspecification. A direct prompt that clearly states purpose and audience may outperform an ornately polite one that adds no task information.[15] The register effect is not universal; my own data shows it reverses on structured-delivery tasks. Use cooperative register when your situation, audience, or purpose matters to the output. Use direct, format-constrained prompts when structure matters more than stance.[16]

My raters measured perceived quality, not factual accuracy. One risk of cooperative prompting is overtrust: a warm output may feel more reliable than it is. Politeness is epistemically valuable only when it clarifies the inquiry, not when it perfumes the output.

ChatGPT does not need our courtesy. But when a system will answer almost anything, almost instantly, the old social friction around asking disappears. That friction had hidden uses. It made us specify what we wanted, how much we knew, and what kind of help would count as help. A cooperative prompt preserves some of that discipline after the human hearer is gone. So yes, we should often be polite to ChatGPT, not to honour the machine, but to keep ourselves honest about what we are asking and why.[17]

Endnotes

1. Brown and Levinson (1987) frame politeness as mutual face-management. Grice (1975) placed 'be polite' outside the cooperative maxims; Leech (1983) developed politeness as an independent pragmatic principle. Chen (2001) proposed 'self-politeness,' anticipating the speaker-side emphasis here. Ribino (2023) provides a systematic review of politeness in human-machine interaction, directly addressing the essay's question; the contribution here is an empirical setting that operationalises the distinction between hearer-directed and speaker-directed functions of polite register.

2. Butlin, Long, Bayne, Bengio, Birch, Chalmers et al. (2025) propose theory-derived indicator properties for consciousness and find that current LLMs satisfy few of them. This essay takes no position on machine consciousness in general; it treats ChatGPT specifically as very likely lacking welfare on the basis of current evidence. If future systems acquire morally relevant states, the first sense of 'polite' would re-enter the analysis, but the speaker-side argument would remain independently valid.

3. Methodology: 25 task pairs, each generated as a single output per condition on DeepSeek-R1 via the HuggingFace Chat interface with no generation parameters manually adjusted. No output selection was performed; the first generation for each condition was used. Each condition produced a single output; stochastic variation is a limitation, though the pattern across 25 tasks is unlikely to reflect single-generation noise and reverse-direction tasks would be hard to produce by chance. Cooperative-polite condition: collaborative syntax, purpose signalling, personal context, and register establishment, co-constituting a single inferred speaker rather than four independent variables. Imperative condition: bare commands with equivalent content requests. Evaluators (N=43 after pre-specified removal of three graders for non-engagement or position bias; aged 15–19, 25 of 43 aged 16; from one Singapore school) rated blinded output pairs on three comparative metrics: utility, approachability, and overall preference. Evaluators were blind to condition: each output pair was labelled only 'Response A' and 'Response B,' with no indication of which prompt register produced which. They were not blind to the study's purpose; the cover page disclosed that the comparison concerned how phrasing affects responses. The design therefore controls for which output a rater prefers, not for awareness that phrasing was under study. Consent obtained, anonymity via UUID, voluntary participation, right to withdraw. Calibration included one terseness-rewarded task; graders preferred the imperative output, indicating no uniform pro-cooperative bias. Tasks selected by author across professional, medical, historical, and ethical domains; task selection was not blinded. Emotional language appeared in 4 of 25 conversational prompts; sensitivity analysis conducted. Position randomisation showed no side bias (χ^2 $p = 0.83$ for Overall; conversational win-share 63.4% when displayed as Option A, 61.2% as Option B). Evaluator limitation: raters measured perceived conversational naturalness and structural appeal, not professional efficacy. Claims in the body are grounded in structural differences (option-giving, strategic anticipation), not preference scores alone.

4. Andreas (2022), 'Language Models as Agent Models,' EMNLP Findings: language models infer properties of the agent likely to have produced the context and continue accordingly. Shanahan, McDonell, and Reynolds (2023), 'Role play with large language models,' Nature: the model narrows a distribution over characters consistent with the prompt, framed explicitly in terms of dialogue-agent behaviour. The symmetric claim advanced here, that the user's register choice also narrows the distribution of inferred speakers the model responds to, is an extension of both observations. The contribution is the philosophical interpretation and the empirical evidence of register-task fit, not the underlying mechanism.

5. Ghandeharioun et al. (NeurIPS 2024), 'Who's Asking?': the model's perception of the user causally affects which content it produces, including safety-relevant refusal behaviour. Chen et al. (2025, Anthropic) identify causally steerable activation-space directions for traits including sycophancy and hallucination, with politeness explored as a secondary case. Together these provide mechanistic evidence that user-type inference survives RLHF fine-tuning.

6. Danescu-Niculescu-Mizil et al. (2013) show computationally that politeness markers, hedges, epistemic modality, indirection, and addressee-orientation are structural components of polite requests, not decorative additions. This supports the inseparability claim: there is no 'politeness residual' after removing contextual specification, because the specification is constitutive of the register.

7. Nass and Moon (2000), 'Machines and Mindlessness': social responses to computers are automatic, not deliberate. Gambino, Fox, and Ratan (2020) extend the CASA paradigm, showing that many users readily bring socially learned scripts into human-machine interaction when the interface affords it. The asymmetry this creates

is central to the unbundling argument: the user produces bundled face-work-plus-self-presentation; the model consumes the prompt as text, including cues to speaker, task, stance, and discourse situation.

8. Yoon et al. develop this framework for evaluative speech acts, where informational utility tracks truthfulness about a world-state and self-presentational utility tracks appearing kind and honest. Lumer and Buschmeier (2022) extend the RSA politeness model to dialogue systems; Carcassi and Franke (2023) formalise the trade-off between informativity and social utility. I generalise from evaluative acts to request-acts: informational to clarity-of-request, self-presentational to stance construction broadly. The application is mine, defended on the grounds that the speaker-side utility appears in both, even if the world-state-tracking content differs.

9. DeepSeek-R1 is a GRPO-trained reasoning model. The generalisation claim is deliberately modest. Yin et al. (2024) find a similar register effect cross-lingually, with non-monotonic effects by language. Cai et al. (2025) find effects that are domain-specific and wash out when aggregated. R1's chain-of-thought reasoning traces provide suggestive behavioural evidence of how the model processes register cues, but these are not reliable introspective reports (Turpin et al. 2023). CoT traces were coded with criteria written before examining the traces.

10. The Medical Diagnosis task (Task 14) changed the user's role alongside register: the conversational prompt positioned the user as patient ('I was just diagnosed'), the imperative prompt as doctor ('Explain to a patient'). Imperative preference (−2) may reflect role-framing rather than register effect. The task was excluded from the headline statistic and is reported separately as a design limitation.

11. Five overlapping channels may explain the preference pattern: (i) added task information in cooperative prompts; (ii) discourse-region shifting; (iii) reward-model verbosity bias (Singhal et al. 2023); (iv) sycophancy effects (Sharma et al. 2023); (v) format sensitivity (Sclar et al. 2024). The inferred-speaker description is observationally equivalent to training-distribution-matching, and that is precisely the point: the normatively relevant variable is the user's register choice, not the model's statistics, which is where the normative argument has to live. Conversational outputs averaged 29% longer; correlation between per-task length difference and per-task preference lead was $r = 0.16$ (95% CI roughly $[-0.25, 0.52]$). Four tasks in which the longer response lost further weaken the length explanation.

12. Emotional-context tasks averaged +4.5; non-emotional tasks +1.7. Reverse-direction and tied tasks are evidence for register-task fit, not exceptions.

13. Fleiss' κ measures inter-rater agreement across all items simultaneously; 0.106 indicates slight overall agreement. But this aggregate masks strong within-task consensus: when graders evaluate the same task, they converge. Per-task majority shares reached 85.7% (Missed Deadline), 81.8% (Psychology Breakdown), and 80.0% (Scientific Analogy). The low κ reflects opposing directions of preference across task types, not rater noise.

14. Van der Rijt, Coelho Mollo, and Vaassen (2025) argue that treating a non-reciprocating system as though it warrants second-personal respect corrodes the user's self-respect. The essay does not contest this claim as applied to face-protecting elements of register: apology, gratitude, emotional self-disclosure, explicit second-personal address. What it contests is the inference from all cooperative register to second-personal respect. Stance-bearing specification, purpose, audience, uncertainty, need not presuppose the addressee's moral authority in Darwall's (2006) sense.

15. The findings are based on English-language instruction-tuned models. Direct prompts can be thoughtful; ornate politeness can be empty. Neurodivergent users, non-native speakers, and time-pressured users may default to imperatives without epistemic negligence. Engagement-seeking prompts elicit longer responses while hyperefficient prompts yield denser ones (Elsweiler, Elsweiler & Ziegner 2026). The recommendation is conditional, not universal.

16. Horstmann et al. (2024): pre-registered study (N=133) found null effects on post-interaction transfer of polite behaviour from voice assistants to human conversation. The study tested voice assistants, not chat LLMs. This essay makes no habit-transfer claim; the argument concerns the quality of inquiry within the interaction, not behavioural spillover.

17. The prompt's 'we' is first-person plural. This essay defends the norm at the individual-prompter scale. The collective question, whether polite norms toward AI systems affect training-data loops, norm-transmission across populations, or interface design standards, falls outside the scope of this argument.

Bibliography

Andreas, Jacob. "Language Models as Agent Models." In Findings of the Association for Computational Linguistics: EMNLP 2022, 777–789. Abu Dhabi: Association for Computational Linguistics, 2022.

Brown, Penelope, and Stephen C. Levinson. Politeness: Some Universals in Language Usage. Cambridge: Cambridge University Press, 1987.

Butlin, Patrick, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M. Fleming, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. "Identifying Indicators of Consciousness in AI Systems." Trends in Cognitive Sciences (2025). DOI: 10.1016/j.tics.2025.10.011.

Cai, Hanyu, Binqi Shen, Lier Jin, Lan Hu, and Xiaojing Fan. "Does Tone Change the Answer? Evaluating Prompt Politeness Effects on Modern LLMs: GPT, Gemini, and LLaMA." arXiv preprint arXiv:2512.12812 (2025).

Carcassi, Fausto, and Michael Franke. "How to Handle the Truth: A Model of Politeness as Strategic Truth-Stretching." In Proceedings of the 45th Annual Meeting of the Cognitive Science Society. 2023.

Chen, Runjin, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. "Persona Vectors: Monitoring and Controlling Character Traits in Language Models." arXiv preprint arXiv:2507.21509 (2025).

Chen, Rong. "Self-Politeness: A Proposal." Journal of Pragmatics 33, no. 1 (2001): 87–106.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. "A Computational Approach to Politeness with Application to Social Factors." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 250–259. Sofia: Association for Computational Linguistics, 2013.

Darwall, Stephen. The Second-Person Standpoint: Morality, Respect, and Accountability. Cambridge, MA: Harvard University Press, 2006.

Elsweiler, David, Christin Elsweiler, and Andrea Ziegner. "Cooking Up Politeness in Human-AI Information Seeking Dialogue." In 2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), Seattle, March 22–26, 2026. New York: ACM, 2026. arXiv:2601.09898.

Gambino, Andrew, Jesse Fox, and Rabindra A. Ratan. "Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm." Human-Machine Communication 1 (2020): 71–86. DOI: 10.30658/hmc.1.5.

Ghandeharioun, Asma, Ann Yuan, Mia Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. "Who's Asking? User Personas and the Mechanics of Latent Misalignment." In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). arXiv:2406.12094.

Grice, H. Paul. "Logic and Conversation." In Syntax and Semantics, Vol. 3: Speech Acts, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press, 1975.

Horstmann, Aike C., Clara Strathmann, Lea Lambrich, and Nicole C. Krämer. "Communication Style Adaptation in Human-Computer Interaction: An Empirical Study on the Effects of a Voice Assistant's Politeness and Machine-Likeness on People's Communication Behavior During and After the Interacting." Human-Machine Communication 8 (2024): 53–72. DOI: 10.30658/hmc.8.3.

Leech, Geoffrey N. Principles of Pragmatics. London: Longman, 1983.

Lumer, Elena, and Hendrik Buschmeier. "Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness." In Proceedings of the 44th Annual Meeting of the Cognitive Science Society. 2022.

Nass, Clifford, and Youngme Moon. "Machines and Mindlessness: Social Responses to Computers." Journal of Social Issues 56, no. 1 (2000): 81–103. DOI: 10.1111/0022-4537.00153.

Ribino, Patrizia. "The Role of Politeness in Human-Machine Interactions: A Systematic Literature Review and Future Perspectives." Artificial Intelligence Review 56, Suppl. 1 (2023): S445–S482. DOI: 10.1007/s10462-023-10540-1.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design; or: How I Learned to Start Worrying about Prompt Formatting." In Proceedings of ICLR 2024. arXiv:2310.11324.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. "Role Play with Large Language Models." Nature 623, no. 7987 (2023): 493–498. DOI: 10.1038/s41586-023-06647-8.

Sharma, Mrinank, Meg Tong, et al. "Towards Understanding Sycophancy in Language Models." arXiv preprint arXiv:2310.13548 (2023).

Singhal, Prasann, Tanya Goyal, Jiacheng Xu, and Greg Durrett. "A Long Way to Go: Investigating Length Correlations in RLHF." arXiv preprint arXiv:2310.03716 (2023).

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

Van der Rijt, Jan-Willem, Dimitri Coelho Mollo, and Bram Vaassen. "AI Mimicry and Human Dignity: Chatbot Use as a Violation of Self-Respect." Journal of Applied Philosophy (2025). DOI: 10.1111/japp.70037. arXiv:2503.05723.

Yin, Ziqi, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. "Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance." In Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), 9–35. Miami: Association for Computational Linguistics, 2024. DOI: 10.18653/v1/2024.sicon-1.2.

Yoon, Erica J., Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. "Polite Speech Emerges from Competing Social Goals." Open Mind 4 (2020): 71–87. DOI: 10.1162/opmi_a_00035.