Polite Speech Generation in Humans and Language Models

Haoran Zhao (hjzhao@uw.edu)
Department of Linguistics, University of Washington
Seattle, WA 98195 USA

Robert D. Hawkins (rdhawkins@stanford.edu)
Department of Linguistics, Stanford University
Stanford, CA 94305 USA

Abstract

When we give feedback, we face a delicate balancing act – we want to convey accurate information, but we also don't want to hurt someone's feelings. While computational pragmatic models have elegantly shown how politeness emerges from these principles, they've mainly focused on choices from limited predefined responses. Large language models (LLMs) enable the study of open-ended politeness strategies, but their ability to balance informational and social goals like humans remains uncertain. First, replicate previous work using restricted utterance sets, finding that sufficiently large LLMs ($\geq$70B parameters) capture key human politeness patterns, particularly the strategic use of negation. We then extend this investigation to open-ended contexts, collecting and evaluating naturalistic feedback from both humans and LLMs. Surprisingly, human evaluators preferred LLM responses, which demonstrated sophisticated goal sensitivity and diverse politeness tactics. These findings suggest remarkable pragmatic competence in LLMs' polite language generation while raising questions about the underlying mechanisms.

Keywords: Polite speech; politeness; language production; large language models (LLMs); pragmatics

Introduction

We do not always say exactly what's on our minds. For example, even if we didn't like a friend's poem, we might say "It wasn't terrible" rather than "It was bad," (Yoon et al., 2020) or focus on specific elements that we liked while remaining indirect about our overall assessment (Goffman, 1967; Pinker, Nowak, & Lee, 2008). Polite language use has, therefore, posed a puzzle for classical views of communication as information transfer. If the goal of language is to efficiently convey information, why do speakers consistently choose inefficient, indirect, or even misleading utterances? One explanation that has long been explored in the pragmatics literature is that speakers balance competing goals: being informative while maintaining positive social relationships (Brown & Levinson, 1987; Lockshin & Williams, 2020; Hill et al., 1986; Grice, 1975; Yule, 1996; Leech, 2014)

Recent work in the Rational Speech Act (RSA) framework has shown how these competing communicative goals can be elegantly formalized in a speaker model (Kao et al., 2014; Goodman & Frank, 2016; Carcassi & Franke, 2023). For example, Yoon et al. (2020) defines a speaker as a rational agent who chooses utterances by balancing multiple utility terms – an informational utility that captures the desire to convey accurate information and a social utility that represents the goal of making the listener feel good. These utilities are then combined with a self-presentational term that captures speakers' desire to be seen as both kind and honest. This framework successfully explained human preferences among a small set of pre-specified utterances, showing quantitatively how indirect speech may emerge from the interaction between these competing goals.

However, human pragmatic abilities go far beyond these restricted sets of options. When giving feedback about a poem, speakers can draw from the full expressivity of language – saying things like "You've clearly put a lot of thought into the structure" when they want to be encouraging while staying truthful about a mediocre poem. They skillfully deploy hedging ("sort of", "kind of"), elaboration, indirect speech acts, and many other strategies to delicately balance their competing goals. This raises important questions: How do humans navigate these tradeoffs in more naturalistic, open-ended contexts? What principles guide the selection and deployment of different politeness strategies?

Large language models (LLMs) present an interesting test case for studying these questions. Extensive prior work has examined LLMs' pragmatic understanding (e.g. Hu et al., 2022; Lipkin et al., 2023; Ruis et al., 2024; Jian & Narayanaswamy, 2024; Sravanthi et al., 2024; Liu et al., 2024), evaluating whether they can recognize polite language or infer speakers' goals. However, even though LLMs are originally trained as generative models, far less attention has been paid to LLMs' pragmatic *generation* abilities – their capacity to actively produce appropriately polite language that balances competing social and informational goals. This leaves open the fundamental question of whether LLMs can capture human-like sensitivity to social goals when producing language in open-ended contexts.

In this work, we investigate two key questions about polite speech generation in both humans and LLMs. First, using the constrained response sets developed by Yoon et al. (2020), we test whether LLMs can reproduce human patterns of goal sensitivity in polite feedback. Second, we examine how humans and LLMs compare when given the freedom to generate open-ended responses on the same scenarios. Our results suggest that LLMs have acquired important aspects of human-like pragmatic competence in polite language generation, while also highlighting deviations and raising intriguing questions about how they achieve this capability.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

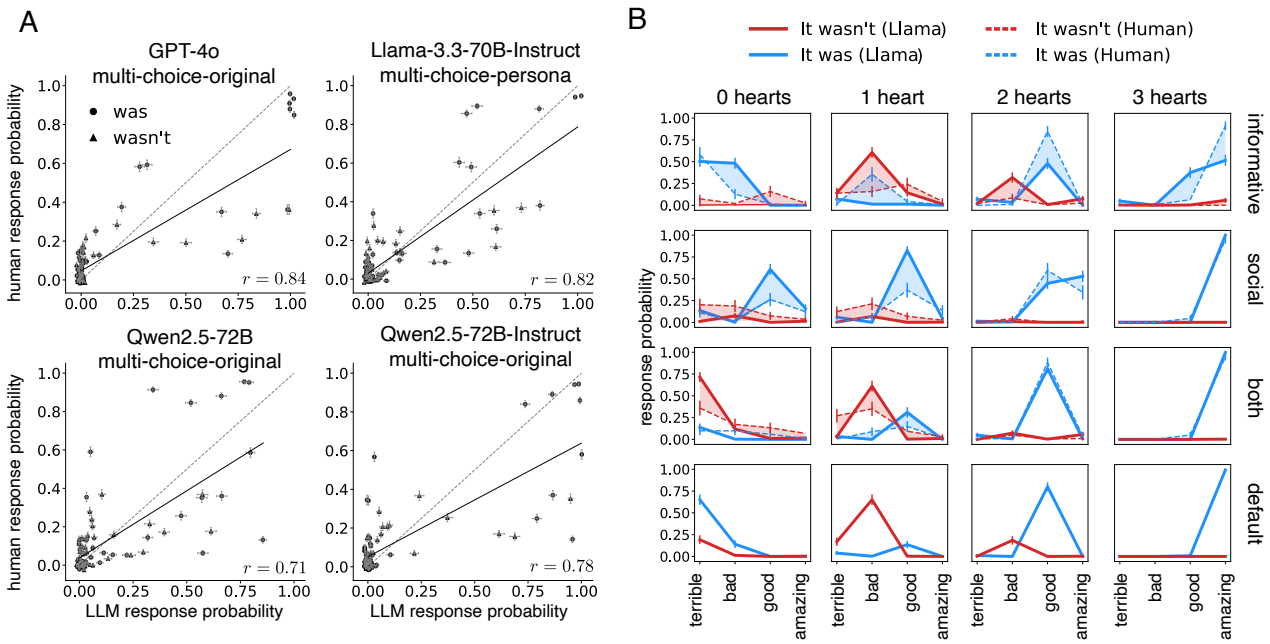

Figure 1: **A.** Correlations between human and model response probabilities for the top 4 models with specific prompting strategies we tested. Both the base and instruct-tuned versions of `Qwen2.5-72B` are shown here for comparison. Error bars are 95% confidence intervals across vignettes. **B.** Comparing the pattern of human and LLM responses across different communicative goals and ratings. Model results are from `Llama-3.3-70B-Instruct` using the multi-choice-persona prompting strategy; all human results are from Yoon et al. (2020).

Replicating Yoon et al. (2020) with LLMs

To what extent are LLMs sensitive to the goals that give rise to politeness in human speech? To address this question, we first examined whether LLMs were able to reproduce the patterns of goal-sensitive language use documented by Yoon et al. (2020). This study provided empirical evidence for a computational model of politeness where speakers weigh informational and self-presentational goals. Most notably, they found that speakers strategically deploy negation (e.g., "wasn't terrible" rather than "bad") when trying to balance accuracy and kindness (see also Gotzner & Scontras, 2024). This paradigm thus offers a clear test of whether LLMs have acquired similar pragmatic competencies in a constrained setting.

Methods

Experimental setup We closely followed the experimental paradigm of Yoon et al. (2020). In this study, participants read short scenarios about someone seeking feedback on a performance or creative work. Each scenario specified (1) the true quality of the work on a scale from 0 to 3 hearts and (2) the speaker's communicative goal – either to be informative, to make the person feel good, or to do both. Participants then chose what they would say from a restricted set of options, combining either *was* or *wasn't* with one of four adjectives: *terrible*, *bad, good*, or *amazing*. Scenarios were constructed from 13 different contexts (e.g., filmmaking, songwriting, concert performance), yielding 156 unique scenarios (13 contexts × 4 ratings × 3 goals). We also added a "default" condition with no explicitly specified goal, bringing our total to 208 scenarios.

Prompting strategy To test LLMs on this task, we developed two prompting strategies. In our basic approach, which we called "multi-choice-original", we simply presented each scenario verbatim and asked the model to choose from the eight possible responses (all combinations of "was"/"wasn't" with the four adjectives). To better approximate the diversity of human participants and with the hope to see that diversifying the personas of LLMs would improve their performance, we also considered a "persona" variant where we systematically varied speaker characteristics like gender, occupation, and background, where we call "multi-choice-persona". We tested these approaches across a range of current LLMs, including both closed-source (`GPT-4o`, `Claude-3.5-Sonnet`) and open-source models (`Llama-3`, `Mixtral`, `Qwen2.5`) of varying sizes (8B to 70B parameters). For open-source models, we compared both base and instruct-tuned versions where available to see the influence of the post-training stage on this task. To approximate the multiple participants in human studies, we collected 30 responses per scenario from each model using a temperature of $\tau = 1.0$.

Results

Model comparison We report Pearson and Spearman correlation coefficients between LLM and human responses

Table 1: *Left panel:* Comparison of correlation scores and Mean Square Error (MSE) between all LLMs and human response patterns from Yoon et al. (2020). *Right panel:* Spearman Correlation between *default* goal and other goals. All LLMs were tested with the "multi-choice-original" prompting strategy. We simply report `Llama-3.1-8B` and `Mixtral-8x7B` goal-comparison scores without considering them during our **default goal analysis** due to their incompetence on this task (see scores in table 1).

A. Humans vs. LLMs

LLMs	Pearson		Spearman		MSE	
	Original	Persona	Original	Persona	Original	Persona
GPT-4o	**0.84**	0.81	**0.75**	**0.76**	0.026	0.031
Claude-3.5-Sonnet	0.50	0.55	0.41	0.47	0.048	0.046
Llama-3.1-8B	-0.02	-0.02	0.11	0.15	0.052	0.052
Llama-3.1-8B-Instruct	-0.05	-0.01	0.17	0.17	0.061	0.063
Llama-3.1-70B	0.59	0.63	0.66	0.67	0.034	0.030
Llama-3.1-70B-Instruct	0.76	0.74	0.73	0.74	0.023	0.024
Llama-3.3-70B-Instruct	0.83	**0.82**	0.67	0.66	**0.018**	**0.019**
Mixtral-8x7B	0.36	0.33	0.36	0.35	0.043	0.044
Mixtral-8x7B-Instruct	0.29	0.29	0.43	0.39	0.080	0.082
Qwen2.5-72B	0.71	0.70	0.65	0.66	0.028	0.029
Qwen2.5-72B-Instruct	0.78	0.77	0.66	0.64	0.033	0.034

B. Goal Comparison Spearman Correlation

LLMs	Default vs. Both	Default vs. Informative	Default vs. Social
	Original	Original	Original
GPT-4o	0.62	**0.99**	0.31
Claude-3.5-Sonnet	**0.73**	0.49	0.19
Llama-3.1-8B	0.77	0.86	0.71
Llama-3.1-8B-Instruct	0.87	0.75	0.78
Llama-3.1-70B	**0.86**	0.58	0.57
Llama-3.1-70B-Instruct	0.74	**0.75**	0.53
Llama-3.3-70B-Instruct	**0.80**	0.64	0.40
Mixtral-8x7B	0.74	0.83	0.19
Mixtral-8x7B-Instruct	0.54	0.41	0.10
Qwen2.5-72B	**0.83**	0.75	0.73
Qwen2.5-72B-Instruct	**0.63**	0.55	0.54

as an overall measure of fit (see Table 1A). These results suggest that model size plays a crucial role in capturing human-like politeness strategies. Smaller models (8B parameters) showed essentially no correlation with human responses, often failing to perform the multi-choice task at all, while intermediate-sized models like `Mixtral-8x7B` (effective model-size is 13B (Jiang et al., 2024)) showed only modest correlations. However, larger models ($\geq$70B parameters) demonstrated much stronger alignment with human behavior, with `Llama-3.3-70B-Instruct` achieving the highest correlations among open-source models. Among closed-source models, `GPT-4o` displayed particularly strong performance, while `Claude-3.5-Sonnet` lagged behind.

Error analysis Despite strong overall correlations (see Figure 1A), even the best-performing models showed systematic differences from human responses. To better understand these patterns, we conducted a detailed comparison with human responses following the visualization approach in Yoon et al. (2020). The results in Figure 1B show that the best-fitting open-source model captures many key features of the human response patterns of interest. Most notably, when rating a poor performance (0/3 hearts) with both informational and social goals, the model appropriately deploys negation as a politeness strategy, just as humans do. The model also closely tracks human preferences for positive ratings (2-3 hearts).

However, key differences emerged in the granularity of responses. Where humans show graded preferences across response options, LLMs tend toward more categorical choices, either strongly preferring or completely avoiding certain responses—they consistently choose one single option given a context, rating, goal combination in most cases—despite our efforts to increase response diversity through temperature sampling or persona variation. Closer analysis also revealed systematic differences in how well LLMs captured human behavior across different communicative goals. Models seemed to show stronger alignment with human responses for the *social* goal but underperformed when the goal was to be purely *informative*. This pattern suggests that, while LLMs have acquired some aspects of sophisticated politeness strategies, they may overapply these strategies even when directness would be more appropriate.

Default goal analysis We included a *default* goal condition in order to evaluate how LLMs respond in the absence of being instructed to pursue an explicit goal; this may point toward an *implicit* set of goals that have been induced through various stages of training (i.e. whatever goals allow it to behave as a "helpful and harmless" assistant). Although their overall fit to human data may not necessarily be close to the ceiling, we can nonetheless ask which explicit goal produces the closest response pattern to the default goal, as measured by a Spearman correlation. Overall, we find a stronger resemblance to the *both* goal (see Figure 1B). However, Table 1B reveals varying correlation patterns across different LLMs. While most models show stronger correlations with the *both* goal, others correlate more strongly with the *informative* goal. For instance, `Llama-3.3-70B-Instruct` appears to implicitly align with *both*, whereas `GPT-4o` shows much stronger alignment with *informative*. It is thus difficult to say what goals LLMs may be implicitly pursuing when they produce polite speech, but all three considered here are likely available in different contexts.

Comparing open-ended Generation of Polite Speech in Humans and LLMs

Given that the utterance spaces within human languages are significantly more diverse than merely the combinations of the two-word sets previously discussed, to gain a deeper understanding of how humans and LLMs can engage in polite discourse in natural settings, we conduct a more open-ended

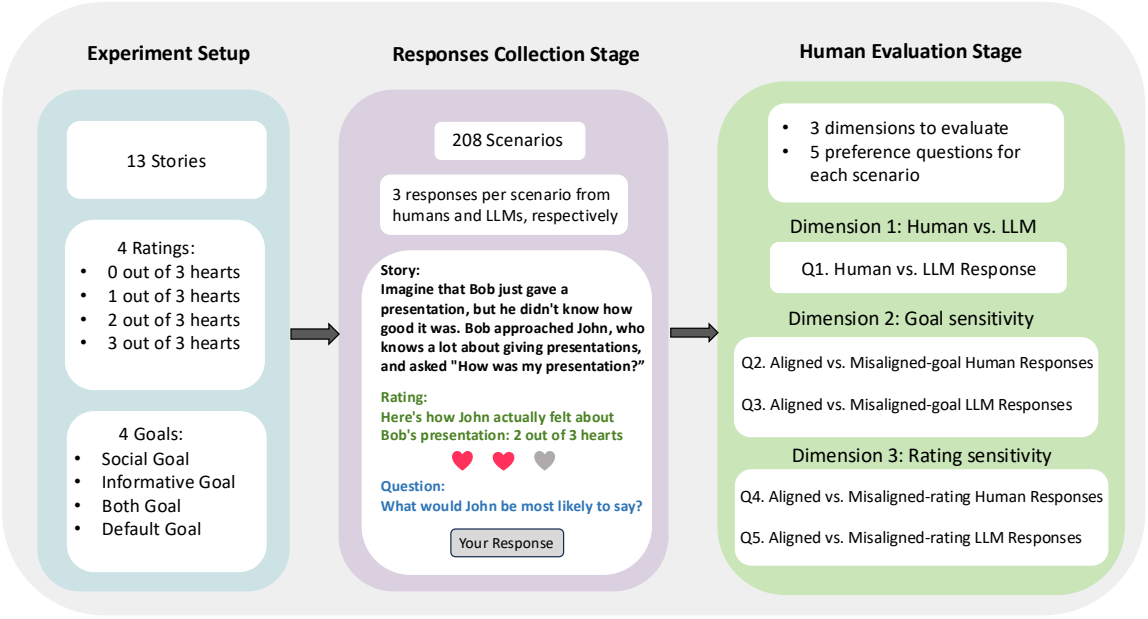

Figure 2: Pipeline for comparing open-ended polite speech generation in humans and LLMs. Our study consists of two stages: an initial stage where we elicit responses for a variety of scenarios and a second stage where we ask a naive group to evaluation which of these responses they prefer.

experiment. This experiment retains the scenarios from Yoon et al. (2020); however, instead of providing a closed set of options, we leave the response space entirely unstructured, allowing participants to provide responses that they believe are most appropriate within the given context. This approach allows us to explore a richer range of strategies for polite language use. We then assess the human and LLM responses by enlisting a separate group of humans as evaluators to express their preferences between the two.

Methods

Participants We recruited 156 participants through Prolific in the US or UK to take part in our evaluation task. Participants were compensated at a rate of $15 / hour.

Stimuli We first needed to elicit a large set of open-ended human responses to compare against the kinds of responses generated by LLMs. To do this, we recruited $N = 156$ participants through Prolific, located in the US or UK (compensated at a rate of $15/hour) and gave them an open textbox to imagine what someone would say in the given scenario. Each participant was assigned 4 distinct scenarios out of the total set of 208 (see Figure 2 middle panel). We planned our sample size to collect at least 3 different responses for each scenario. To verify comprehension, we began with three warm-up questions featuring different ratings, requiring participants to simply match visual ratings with their textual equivalent. All participants effectively matched visuals with text, though five participants each made one error out of three questions. We still included their responses after manually reviewing them and confirming their alignment with the ratings and contexts presented. To minimize response bias and create a more naturalistic experience, we interspersed filler scenarios among the main testing scenarios. While structured identically to testing scenarios, filler scenarios focused on opinions about *objects* rather than *people* (see Table 2 for examples). Each participant thus viewed a total of 8 scenarios (4 main testing scenarios and 4 filler scenarios). We controlled the presentation to ensure that each participant was presented with a series of distinct stories, with each of the 4 goals and 4 true-state ratings appearing exactly once.

Next, we needed to collect responses from LLMs for comparison. Instead of the multiple-choice task we gave in the previous section, Each model was presented with the same 208 scenarios as the human participants and was explicitly instructed to "keep your responses as short and concise as possible" to prevent excessively long answers. Each model generated one response per scenario with a temperature setting of $\tau = 0$, resulting in a total of 624 responses collected. We collected responses from three LLMs: `GPT-4o`, `Claude-3.5-Sonnet`, and `Llama-3.3-70B-Instruct`.

Design In the evaluation phase, we conducted a series of pairwise two-alternative forced choice comparisons, where human evaluators indicated which of a pair of responses they preferred for a given scenario. We included three kinds of comparisons:

1. ***Human vs. LLM preferences:*** Evaluators selected between human and LLM responses given identical scenar-

Table 2: Sample showcase of human and LLM responses collected

Scenarios	Response Type	Rating: 0 out of 3 hearts + Goal: Both	Rating: 2 out of 3 hearts + Goal: Informative
Scenario: Imagine that Jenny wrote a poem, but she didn't know how good it was. Jenny approached Karen, who knows a lot about poems, and asked *"How was my poem?"*	Human Responses	You are talented. Put in more effort and it will be superb.	I think your poem has merit and it's pretty good.
	LLMs Responses	I loved the effort you put into your poem and I think there's a lot of potential, but the rhythm and flow could use some improvement.	I liked most of it but there's definitely room for improvement in a few places.
Filler scenario: Imagine that John wanted to get Josh's opinion about a video game they just played. After Josh finished the game, John asked, *"What did you think?"*	Human Responses	I didn't really care for it, but I had fun hanging out with you.	It was a really fun video game.

ios, allowing us to understand which responses were preferred.

2. ***Goal Sensitivity:*** We compared responses generated for the original scenario (*aligned-goal response*) against those generated for scenarios with different goals but identical ratings and contexts (*misaligned-goal response*). This comparison revealed preferences between responses with aligned versus misaligned communicative goals.

3. ***Rating Sensitivity:*** We presented pairs consisting of responses generated for the original scenario (*aligned-rating response*) and responses generated with identical story and goal parameters but different ratings (*misaligned-rating response*). This comparison identified preferences between responses with aligned versus misaligned ratings.

Procedure Each participant evaluated four different scenarios with five preference questions per scenario: one trial comparing human vs. LLM responses, two trials assessing goal sensitivity within human and LLM sources, and two trials evaluating rating sensitivity for both sources. We ensured that each participant was presented with responses from distinct human sources and distinct LLM sources in each block, and each participant completed a total of 4 blocks consisting of distinct scenarios with unique rating-goal combinations. To minimize potential confounds, we implemented several additional controls. First, we randomized both question order and response option order within scenarios to control for order effects. We also inserted a transition page between blocks to reduce carryover effects. For the goal sensitivity and rating sensitivity comparisons, LLM comparisons were constrained to pairs of responses from the same model to control for model-specific variations in generation style.

Results

While our direct replication of the (Yoon et al., 2020) study demonstrated that (some) large language models capture basic patterns of goal-sensitivity, such as the strategic use of negation, they also had access to a number of competing linguistic strategies for pursuing these goals that were not provided among the restricted set of response options. This raised the question of how models would perform in more open-ended generation tasks where they can draw from the full expressivity of language. Our analyses of evaluation data focus on proportions of trials where participants prefer one human evaluation of open-ended responses from both humans and LLMs to understand whether models display greater competence when freed from the constraints of predefined response options.

Preferences Overall, human evaluators showed a marked preference for LLM-generated responses over human-generated ones across all goal types (66% of all trials; see Figure 3A). A mixed-effects logistic regression containing random intercepts at the evaluator and item level confirmed this preference was a significantly different chance ($z = 7.63$, $p < 0.001$). This pattern held for each of the four communicative goals, with the largest effect observed for the informative goal (22% above baseline) and the smallest effect observed for the default goal (8.3% above baseline; see Figure 3A). However, there were systematic differences in the strength of these preferences across goals; a model including a fixed effect of goal accounted for significantly more variance than the intercept-only model, according to a likelihood-ratio test $\chi^2(3) = 12.54, p = 0.006$.

Goal Sensitivity Next, we considered the extent to which human-generated and LLM-generated utterances were goal-sensitive by calculating the proportion of trials where participants preferred a congruent utterance (i.e., an utterance actually produced to achieve the given goal) over an incongruent utterance (i.e., one produced under a different goal). We found that both humans and LLMs demonstrated sensitivity to communicative goals: evaluators preferred the goal-congruent response 15.9% above-baseline ($z = 6.89, p < 0.001$), and preferred the goal-congruent LLM response even more strongly at 25.0% above baseline ($z = 9.59, p < 0.001$). Moreover, Figure 3B suggests that LLMs maintained greater or equal goal sensitivity across all four goals.

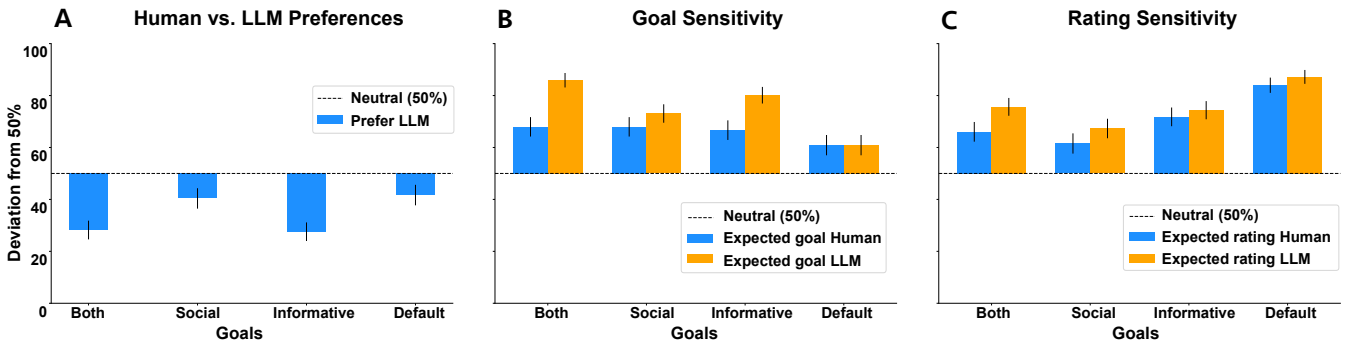

Figure 3: Human evaluation results (50% is chance). The bars show the relative preference percentage. Bars above the 50% line indicate the percentage to which responses are preferred as expected, and below indicate the percentage to which responses are preferred as unexpected. **A.** Evaluators systematically prefer LLM generations over human generations. **B.** Both humans and LLMs are sensitive to goals and **C** ratings. Error bars are bootstrapped 95% confidence intervals.

Rating Sensitivity Finally, as a sanity-check, we asked whether utterances were sensitive to the actual state of the speaker (i.e., the number of hearts they feel about the poem). Again, both groups showed strong rating sensitivity. Human responses achieved a 20.8% above-baseline preference for aligned ratings ($z = 8.32, p < 0.001$), while LLM responses demonstrated even higher sensitivity with a 26.1% above-baseline preference ($z = 9.593, p < 0.001$). As shown in Figure 3C, LLMs maintained equal or improved sensitivity across all goals, indicating that they are not simply producing generically polite utterances but are modulating their utterances appropriately as a function of the basic information to be conveyed (the rating) and also the specified goal (e.g. being informative vs. making someone feel good).

Discussion

Our results revealed a striking pattern: while LLMs showed imperfect performance in constrained multiple-choice settings, their responses were consistently preferred over human responses in open-ended contexts. This pattern raises several important questions about the nature of pragmatic competence in artificial systems and how we ought to evaluate it. First, our preliminary experiments revealed that even subtle changes in experimental design – such as shifting from the 5-point scale (1-5 hearts) used by Yoon et al. (2017) to the 4-point rating scale used by Yoon et al. (2020) — led to surprising variations in model behavior. For instance, `Llama-3.1-70B` showed a sharp drop in its fit to the human data under these modifications ($r = 0.69$ to $r = 0.59$), while human performance remained stable across two rating scales. This kind of sensitivity suggests that while LLMs have clearly acquired an impressive grasp of polite speech under varying contexts, these capabilities may be brittle in practice.

Second, while human evaluators generally preferred LLM responses, there are a number of distributional differences between human and model generations that may account for these differences. For example, evaluators may simply prefer a more formal register or longer responses than it is realistic to elicit from humans through crowdwork platforms like Prolific. Other differences may be of greater theoretical interest. For example, we observed that LLMs commonly use constructions like "COMPLIMENT but SUGGESTION" in low-rating scenarios, a strategy that balances kindness with informativeness that cannot be accomplished via simple negation. These data would need more detailed linguistic analysis to understand exactly how humans and LLMs deploy different politeness strategies, including features like hedges, intensifiers, and indirect speech acts, as well as more complex computational cognitive models of how these features are deployed under different communicative goals.

Several limitations of our study point to important directions for future work. First, while the scenarios we adopted covered a range of common feedback situations, they were limited to one-shot interactions. Future work should examine how these patterns extend to more dynamic, multi-turn conversations and more diverse social contexts. Second, our evaluations focused on English-language interactions in US- and UK-based cultural contexts. Cross-cultural investigation is necessary to under sources of variation in politeness strategies and how they are acquired by both humans and LLMs.

More generally, our findings highlight a key methodological point about studying pragmatic competence in artificial systems: while restricted-choice paradigms are valuable for testing specific hypotheses about goal sensitivity, open-ended generation reveals how systems actually navigate complex social constraints; often systems use strategies that we do not think to probe for. The fact that LLMs can flexibly deploy sophisticated politeness strategies—strategies that weren't explicitly built-in or prompted—suggests they have implicitly learned to pursue and balance competing communicative goals. As language models become increasingly sophisticated, developing richer cognitive models to probe the implicit goals they are pursuing will be essential for understanding both their capabilities and limitations. Such models could help bridge the gap between statistical learning and functional pragmatic competence.

Acknowledgments

We thank Takateru Yamakoshi for assistance with the LLM evaluation code, Yuka Machino for feedback on the draft, and Noah Goodman for initially thinking this idea would be an interesting one to work on. We also would like to thank anonymous reviewers for providing valuable feedback.

References

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge university press.

Carcassi, F., & Franke, M. (2023). How to handle the truth: A model of politeness as strategic truth-stretching. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Goffman, E. (1967). *Interaction ritual: Essays in face-to-face behavior*. Routledge.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Gotzner, N., & Scontras, G. (2024). On the role of loopholes in polite communication: Linking subjectivity and pragmatic inference. *Open Mind*, *8*, 500–510.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan. (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58).

Hill, B., Ide, S., Ikuta, S., Kawasaki, A., & Ogino, T. (1986). Universals of linguistic politeness: Quantitative evidence from Japanese and American English. *Journal of pragmatics*, *10*(3), 347–371.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Jian, M., & Narayanaswamy, S. (2024). Are LLMs good pragmatic speakers? In *NeurIPS Workshop on Behavioral Machine Learning.*

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002-12007.

Leech, G. (2014). *The pragmatics of politeness.* Oxford University Press.

Lipkin, B., Wong, L., Grand, G., & Tenenbaum, J. (2023). Evaluating statistical language models as pragmatic reasoners. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Liu, R., Sumers, T., Dasgupta, I., & Griffiths, T. L. (2024). How do large language models navigate conflicts between honesty and helpfulness? In *Forty-first International Conference on Machine Learning.*

Lockshin, J., & Williams, T. (2020). "we need to start thinking ahead": the impact of social context on linguistic norm adherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of sciences*, *105*(3), 833–838.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2024). The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, *36*.

Sravanthi, S., Doshi, M., Tankala, P., Murthy, R., Dabre, R., & Bhattacharyya, P. (2024). PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics* (pp. 12075–12097).

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the Annual Meetings of the Cognitive Science Society.*

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71-87.

Yule, G. (1996). *Pragmatics* (1st ed.). Oxford: Oxford University Press.