Comparing human and LLM politeness strategies in free production

Haoran Zhao
Department of Linguistics
University of Washington
hjzhao@uw.edu

Robert D. Hawkins
Department of Linguistics
Stanford University
rdhawkins@stanford.edu

Abstract

Polite speech poses a fundamental alignment challenge for large language models (LLMs). Humans deploy a rich repertoire of linguistic strategies to balance informational and social goals – from positive approaches that build rapport (compliments, expressions of interest) to negative strategies that minimize imposition (hedging, indirectness). We investigate whether LLMs employ a similarly context-sensitive repertoire by comparing human and LLM responses to English-language scenarios in both constrained and open-ended production tasks. We find that larger models ($\geq$70B parameters) successfully replicate key effects from the computational pragmatics literature, and human evaluators prefer LLM-generated responses in open-ended contexts. However, further linguistic analyses reveal that models disproportionately rely on negative politeness strategies to create distance even in positive contexts, potentially leading to misinterpretations. While LLMs thus demonstrate an impressive command of politeness strategies, these systematic differences provide important groundwork for making intentional choices about pragmatic behavior in human-AI communication.

1 Introduction

Speakers do not always say exactly what they mean. For example, we might say a friend's poem "wasn't terrible" rather than saying "it was bad" to avoid hurting their feelings (Yoon et al., 2020), or just compliment specific elements that we liked without mentioning other elements we didn't like (Goffman, 1967; Pinker et al., 2008). These kinds of politeness strategies allow speakers to balance competing goals, conveying accurate information while maintaining positive relationships (Hill et al., 1986; Leech, 2014). As large language models (LLMs) are increasingly deployed in open-ended interactions across sensitive social domains like healthcare and education, their ability to appropriately use and understand polite language remains an important alignment challenge.

Politeness theory provides a valuable framework for addressing these questions. Seminal work by Brown and Levinson (1987) distinguishes between positive politeness strategies that affirm the listener (compliments, expressions of interest) and negative politeness strategies that minimize imposition (hedging, indirectness). While subsequent work has expanded this framework to encompass broader relational and rapport management concerns (Spencer-Oatey, 2011; Watts, 2003; Locher, 2013), this basic distinction remains crucial: different contexts call for different strategies, and mismatches can lead to communication breakdowns. If an AI system employs negative hedging strategies ("I am somewhat concerned that this approach might not be optimal") in contexts where human speakers would expect positive, rapport-building strategies ("I love your creativity here, and wonder if we could build on it by..."), users may be left unsure whether the hedging reflects a genuinely negative evaluation or simply a systematic bias in expressions of politeness.

This kind of pragmatic misalignment represents a critical gap in our understanding of LLMs as social agents. While considerable attention has been paid to whether models can recognize politeness or generate polite language in constrained settings, effective social interaction depends not just on understanding politeness norms in the abstract but on actively selecting and applying appropriate strategies from a diverse linguistic repertoire. Human speakers navigate this complexity intuitively, deploying hedging, elaboration, indirect speech acts, and numerous other strategies to balance competing communicative goals in context-sensitive ways. To fully understand LLMs' grasp of politeness strategies, we need to examine whether they exhibit similar patterns of strategy selection and deployment across different contexts. This requires

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16200–16228
November 4-9, 2025

moving beyond limited-choice evaluations to examine open-ended language generation, where models have access to the full range of linguistic choices. Rather than prescribing how AI systems ought to handle politeness, we seek descriptive evidence about current patterns of behavior.

In this work, we investigate we compare human and LLM politeness strategies in English-language scenarios, making the following contributions:

- We test whether LLMs reproduce human patterns of goal sensitivity in polite feedback using constrained response sets from Yoon et al. (2020).
- We collect and analyze a new dataset of open-ended responses from both humans and LLMs to identical social scenarios, enabling direct comparison of politeness strategies.
- We perform detailed linguistic analyses to identify systematic differences in how humans and LLMs deploy various categories of politeness strategies.

Our results reveal that while LLMs have acquired important aspects of human-like pragmatic competence in polite language production – enough to be preferred by human evaluators – they also show systematic differences in strategy deployment that raise intriguing questions about the mechanisms underlying their social language capabilities. In particular, we find that models disproportionately rely on negative politeness strategies (minimizing imposition) even in contexts where humans prefer positive politeness strategies (building rapport), suggesting important differences in how these systems navigate social interactions[1].

2 Related Work

2.1 Computational models of politeness

Research on politeness in linguistics and cognitive science has evolved from descriptive frameworks to quantitative models of pragmatic language use. Foundational work by Brown and Levinson (1987) established a systematic taxonomy of strategies, which has provided conceptual scaffolding for subsequent computational approaches. Studies in computational linguistics have since documented various linguistic markers of politeness across languages and contexts, examining formal features such as hedging, indirectness, and specific syntactic constructions correlate with perceived politeness (Danescu-Niculescu-Mizil et al., 2013; Aubakirova and Bansal, 2016).

Recent models in the Rational Speech Act (RSA) framework have explained the use of polite language as emerging from tradeoffs between informational utility capturing the desire for accuracy, a social utility representing the goal of making listeners feel good, and a self-presentational term reflecting speakers' desire to be perceived as both kind and honest (Yoon et al., 2020; Lumer and Buschmeier, 2022; Carcassi and Franke, 2023; Gotzner and Scontras, 2024). This body of work has established a solid theoretical foundation for analyzing politeness as a pragmatic phenomenon arising from underlying tradeoffs. However, existing models have primarily focused on explaining choices among a small number of constrained utterance alternatives rather than modeling the rich variety of strategies humans employ in open-ended generation contexts.

2.2 Pragmatic capabilities in LLMs

Recent research has explored various aspects of pragmatic competence in large language models. Studies have examined LLMs' ability to understand indirect speech acts (Ruis et al., 2024; Jian and Narayanaswamy, 2024), recognize conversational implicatures (Hu et al., 2022; Lipkin et al., 2023), and interpret non-literal language (Yerukola et al., 2024; Liu et al., 2024). These investigations predominantly employ multiple-choice formats, presenting models with pragmatic puzzles and evaluating their ability to select contextually appropriate interpretations. Results generally suggest that modern LLMs demonstrate sophisticated pragmatic understanding, often approaching human-like performance on benchmark tasks. However, these studies primarily assess recognition rather than production capabilities, leaving open questions about whether models can actively deploy pragmatic strategies in their own generated outputs.

2.3 Polite language generation in LLMs

Work on generating polite language in AI systems represents a small but growing research area. Early approaches focused on style transfer, with systems like those developed by Niu and Bansal (2018) demonstrating that neural models could transform neutral text into more polite versions through specific syntactic transformations. Subsequent work explored paraphrasing to increase po-

[1]Data & Code: https://github.com/haoranzhao419/politeness-speech-production

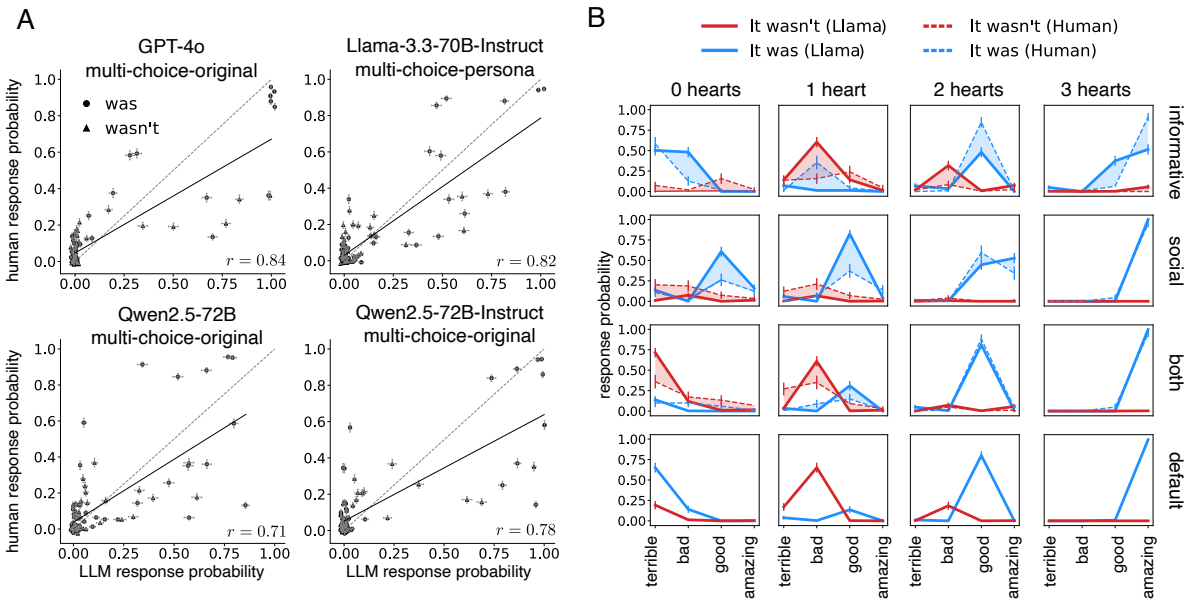

Figure 1: (A) Correlations between human and model response probabilities for the top 4 models with specific prompting strategies we tested. Both the base and instruct-tuned versions of `Qwen2.5-72B` are shown here for comparison. Error bars are 95% confidence intervals across vignettes. (B) Comparing 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; human responses are from Yoon et al. (2020).

liteness (Fu et al., 2020), politeness-focused style transfer (Madaan et al., 2020), and creating polite chatbots (Mukherjee et al., 2023). However, these systems typically focused on surface-level transformations rather than strategic deployment of politeness based on contextual factors. As noted in a recent survey (Priya et al., 2024), existing approaches to polite language generation have predominantly emphasized isolated features (hedging expressions, please markers, specific lexical choices) rather than examining the full repertoire of politeness strategies and how they're selected based on communicative context. This leaves a significant gap in our understanding of whether LLMs can approximate the context-sensitivity that characterizes human politeness. Our work addresses this gap by directly comparing politeness strategies in humans and LLMs across varying communicative goals, examining whether models align with human preferences for positive versus negative politeness strategies in different contexts.

3 Experiment 1: Constrained Settings

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 could reproduce the patterns of goal-sensitive language use reported by Yoon et al. (2020). Their study provided empirical evidence for a computational model of politeness where speakers strategically balance informational accuracy with social goals. Most notably, they found that when giving negative feedback, humans often deploy indirectness through negation (e.g., "wasn't terrible" rather than "bad"; see also Gotzner and Scontras, 2024; Lumer and Buschmeier, 2022).

We reimplemented this experiment with LLMs to assess their pragmatic competence in a constrained setting. In each scenario, a character gives feedback about another character's performance (e.g., a piano play or presentation), with the true quality ranging from 0 to 3 hearts. The speaker has one of four communicative goals: to be *informative*, to be *kind*, or *both*. We also added a *default* condition with no explicit goal specified to understand how LLMs behave by default. Models selected from the same set of eight responses used by humans, combining either "was" or "wasn't" with four adjectives (terrible, bad, good, amazing).

We tested a range of open-source (8B-72B parameters) and closed-source models using two prompting strategies: an "original" strategy that presented scenarios verbatim, and a "persona" variant that systematically varied speaker characteristics (e.g., gender, occupation, background) to better

LLMs	Comparison with humans				Comparison with default goal		
	Spearman		MSE		vs. Both	vs. Inf.	vs. Social
	Original	Persona	Original	Persona			
GPT-4o	**0.75**	**0.76**	0.026	0.031	0.62	**0.99**	0.31
Claude-3.5-Sonnet	0.41	0.47	0.048	0.046	**0.73**	0.49	0.19
Llama-3.1-8B	0.11	0.15	0.052	0.052	0.77	**0.86**	0.71
Llama-3.1-8B-Instruct	0.17	0.17	0.061	0.063	**0.87**	0.75	0.78
Llama-3.1-70B	0.66	0.67	0.034	0.030	**0.86**	0.58	0.57
Llama-3.1-70B-Instruct	0.73	0.74	0.023	0.024	0.74	**0.75**	0.53
Llama-3.3-70B-Instruct	0.67	0.66	**0.018**	**0.019**	**0.80**	0.64	0.40
Mixtral-8x7B	0.36	0.35	0.043	0.044	0.74	**0.83**	0.19
Mixtral-8x7B-Instruct	0.43	0.39	0.080	0.082	**0.54**	0.41	0.10
Qwen2.5-72B	0.65	0.66	0.028	0.029	**0.83**	0.75	0.73
Qwen2.5-72B-Instruct	0.66	0.64	0.033	0.034	**0.63**	0.55	0.54

Table 1: Spearman correlations between the frequencies of human responses (Yoon et al., 2020) and LLM responses across all goal-rating combinations. Bold values indicate the highest correlation.

approximate the diversity in the population of human participants (Murthy et al., 2025; He-Yueya et al., 2024). For each model and prompting strategy, we sampled 30 responses with temperature $\tau = 1.0$ per scenario (see Appendix A for details of prompt design).

3.1 Model comparison

We report Spearman correlation and mean squared error (MSE) between LLM and human responses as an overall measure of fit (see Table 1). These results suggest that model size plays a crucial role in capturing human-like politeness strategies. Smaller models (`Llama-3.1-8B`) 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 (Spearman $r = 0.67$). Among closed-source models, `GPT-4o` displayed particularly strong performance (Spearman $r = 0.75$), while `Claude-3.5-Sonnet` lagged behind with more modest correlations ($r = 0.41$). These findings suggest that sophisticated pragmatic competence for politeness emerges primarily in larger models, potentially reflecting the greater contextual sensitivity needed to balance competing communicative goals.

3.2 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. 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: both humans and LLMs prefer to say "wasn't terrible" rather than "was bad". The model also closely tracks human preferences for positive ratings (2-3 hearts), showing appropriate sensitivity to the quality of the performance.

However, key differences emerged in the granularity of responses. Where humans show graded preferences across response options (distributing probability mass across multiple choices), LLMs tend toward more categorical binary choices, either strongly preferring or completely avoiding certain responses. They consistently choose one single option given a context, rating, and goal combination in most cases—despite our efforts to increase

response diversity through temperature sampling ($\tau = 1.0$) or persona variation in prompting.

Closer analysis also revealed systematic differences in how well LLMs captured human behavior across different communicative goals. Models showed stronger alignment with human responses for the *social* goal but underperformed when the goal was to be purely *informative*. For example, when humans prioritize being informative about a poor performance, they often select direct negative feedback ("was bad"), while LLMs sometimes persist with softened language. 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, potentially reflecting their training to be generally "helpful and harmless".

3.3 Default goal analysis

We included a *default* goal condition (no explicitly specified goal) to evaluate how LLMs respond without specific communicative instructions. This condition helps reveal the implicit goals that might have been induced through various stages of model training. Although the overall fit to human data varies across models, we can ask which explicit goal produces the closest response pattern to the default goal, as measured by Spearman correlation.

Overall, we find a stronger resemblance to the *both* goal (see Figure 1B), suggesting that models generally attempt to balance informativeness and social considerations by default. However, Table 1 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* (Spearman $r = 0.80$), whereas `GPT-4o` shows much stronger alignment with *informative* (Spearman $r = 0.99$).

These varied patterns suggest that the implicit goals guiding different LLMs' polite speech may reflect differences in their training objectives and alignment procedures. The dominant pattern of alignment with the *both* goal is consistent with the general instruction to models to be both helpful (informative) and harmless (socially appropriate). However, the variability across models indicates that while these systems have acquired sophisticated politeness capabilities, the specific ways they balance competing goals may differ from model to model.

4 Experiment 2: Open-ended Generation

While our multiple-choice experiment demonstrated that larger LLMs can reproduce basic patterns of goal-sensitive politeness strategies, such as the strategic use of negation, this constrained format limits our understanding of how models deploy politeness in naturalistic settings. In real-world interactions, speakers draw from a rich repertoire of linguistic devices beyond those provided in fixed-choice scenarios. This raises a critical question: how do LLMs perform when given the freedom to generate polite language from scratch?

To address this question, we designed an open-ended generation experiment that uses the same scenarios as our multiple-choice study but removes the response constraints. This approach allows us to examine whether LLMs employ a similarly diverse and context-sensitive set of politeness strategies as humans when both have access to the full expressivity of language, and directly compare to results in the constrained setting.

4.1 Methods

We used the scenarios from Yoon et al. (2020), preserving the same performance ratings (0-3 hearts) and communicative goals (informative, social, both, default). We collected 3 open-ended responses per scenario from 156 human participants via Prolific (each responding to 4 distinct scenarios) and three responses from LLMs that performed well in our first experiment: `GPT-4o`, `Claude-3.5-Sonnet`, and `Llama-3.3-70B-Instruct`. We selected these larger models after calibration experiments revealed that smaller models failed to generate coherent responses in the open-ended format (see Appendix B.6). Models were instructed to "keep responses short and concise" to ensure comparable length with human responses.

To assess preferences for these responses, we then conducted a two-alternative forced-choice evaluation with 156 human evaluators, each viewing four different scenarios. Evaluators made five judgments per scenario: (1) comparing human vs. LLM responses, (2-3) comparing goal-congruent vs. goal-incongruent responses for both sources, and (4-5) comparing rating-congruent vs. rating-incongruent responses for both sources. We randomized presentation order and ensured evaluators saw responses from different sources across blocks (see Appendix B for full details).

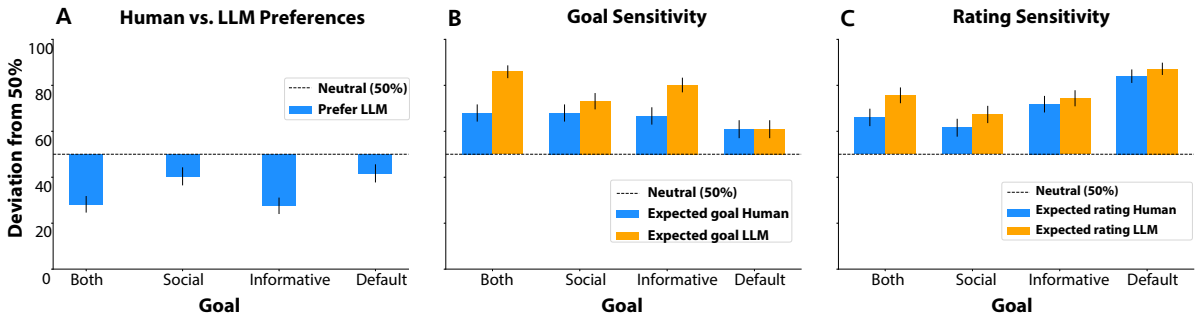

Figure 2: Human evaluation results. The bars show the relative preference (50% is chance). Bars above the 50% line indicate the percentage to which responses are preferred as expected, and below indicate the percentage to which responses are preferred as unexpected. (A) Evaluators systematically prefer LLM generations over human generations. (B) Both humans and LLMs are sensitive to goals and (C) ratings. Error bars are bootstrapped 95% confidence intervals.

4.2 Results

Overall preferences Surprisingly, human evaluators showed a marked preference for LLM-generated responses over human-generated ones across all goal types (66% of all trials; see Figure 2A). A mixed-effects logistic regression containing random intercepts at the evaluator and item level confirmed this preference was significantly different from 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 2A). 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 human 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 2B suggests that LLMs maintained greater or equal goal sensitivity across all four goals, indicating they successfully tailored their language to the specified communicative objective.

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 felt about the performance being evaluated). 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 ($z = 9.59, p < 0.001$). As shown in Figure 2C, LLMs maintained equal or improved sensitivity across all goals, indicating that they are not simply producing generically polite utterances but are modulating their responses appropriately as a function of both the basic information to be conveyed (the rating) and the specified communicative goal (e.g., being informative vs. making someone feel good).

5 Linguistic Analysis of Politeness

While our evaluations show that LLMs successfully generate polite language that human evaluators prefer, these preferences alone don't reveal whether models use the same linguistic mechanisms as humans. To understand the specific politeness strategies employed by both humans and LLMs, we conducted a detailed linguistic analysis of the open-ended responses.

5.1 Negation

As a first step in our analysis, we examined how frequently the strategic use of negation documented in Yoon et al. (2020) and tested in Experiment 1

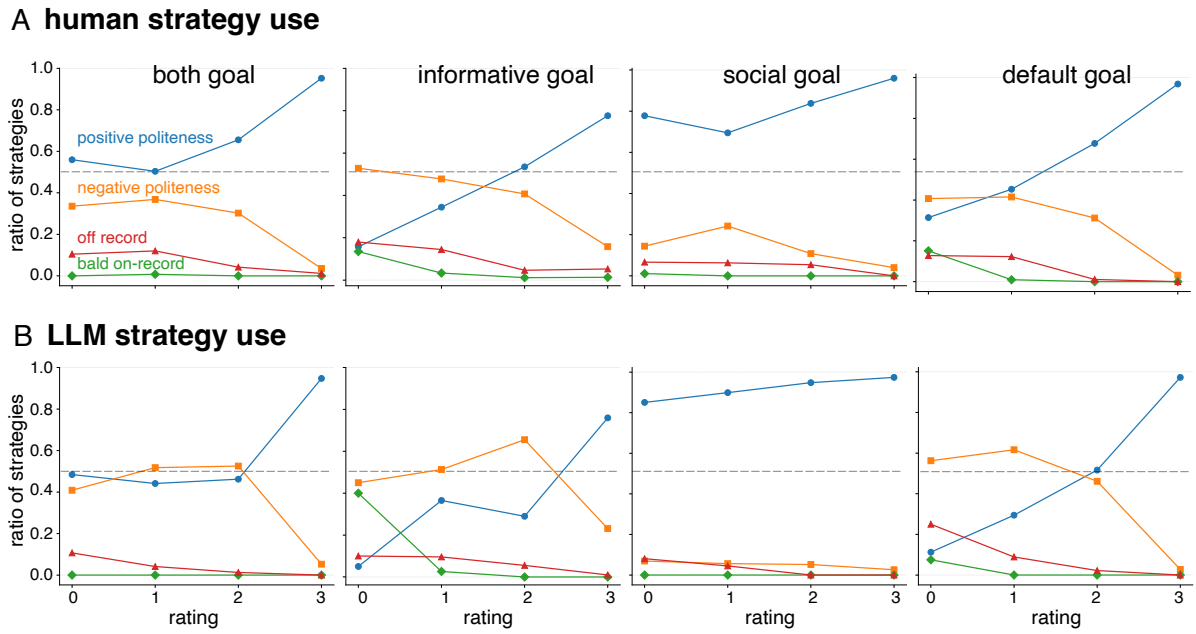

Figure 3: Proportion of different politeness strategies across ratings and goals for (A) human and (B) LLMs.

is employed in open-ended responses. Among all 1,248 responses collected, 527 (42.2%) used the specific pattern of adjective evaluation studied by Yoon et al. (2020). Within this subset, 35 responses (6.6%) employed negation as a politeness strategy, and negation was most common in low-rating (0 or 1 heart) scenarios, which qualitatively replicates our findings from Experiment 1 (see Figure 5 in Appendix for details). Thus, the negation strategies studied in constrained settings do appear in open-ended production, but represent just one of many politeness devices available to speakers.

5.2 Word usage patterns

To better understand differences between human and LLM responses, we analyzed unigram distributions using Pointwise Mutual Information (PMI) and Jensen-Shannon Divergence (JSD). First, examining unigrams with the highest PMI, we found that human responses more frequently incorporated casual language and expressions (e.g., “awesome," “great"), whereas LLM-generated responses tended toward more formal linguistic choices (e.g., “fabulous," “excellent"). Both groups effectively employed personalization as a politeness strategy, such as directly mentioning the performer’s name (e.g., “Your app is pretty good, Henry!"). Additionally, both humans and LLMs adapted their lexical choices based on context, with minimal overlap in high-PMI words across different goals and ratings.

Next, we quantified differences in empirical word frequency distributions by calculating the Jensen-Shannon Divergence (JSD). Interestingly, the JSD between the lexical distributions of preferred and non-preferred response groups was quite small (JSD = 0.013) though still significantly different than a permuted null distribution ($p < 0.001$), while all other group comparisons showed much larger differences (JSD > 0.13, $p < 0.001$; see Table 10). This suggests that simple lexical choice may not be the primary driver of human preferences in polite language.

Finally, we conducted higher-dimensional analyses using SBERT embeddings (Reimers and Gurevych, 2019) to distinguish between response categories. These analyses (described in Appendix C.2) revealed that while human vs. LLM responses were readily distinguishable in embedding space (83% accuracy), preferred vs. non-preferred responses were much harder to classify (54% accuracy). LLM responses were more distinguishable across different communicative goals than human responses, suggesting more stereotyped strategies.

5.3 Annotated politeness strategies

To obtain a comprehensive picture of the politeness strategies employed in human and LLM responses, we conducted a detailed annotation using the politeness framework from Brown and Levinson (1987), supplemented by markers from Danescu-Niculescu-Mizil et al. (2013). This framework distinguishes four broad categories of polite-

Category	Politeness Strategy	Examples
Positive Politeness	**Gratitude**	I'm so *grateful* that . . .
	Be optimistic	You can even make them better next time!
Negative Politeness	**Apologizing**	I didn't like it, sorry!
	Question or hedge	Maybe you can try something different?
Off-Record	**Be vague**	It was *interesting*. (true rating is 0)
	Give association clues	It was better than those who can't play.
Bald on-Record	**Negative lexicon**	It was *terrible*!
	Factuality	I didn't like the cookies at all.

Table 2: Examples of utterances for different politeness strategies (two per category), with the corresponding sub-strategy highlighted in bold. See the comprehensive list in Table 12.

ness strategies with many subtypes (see Table 2 and Appendix Table 12): positive politeness (e.g. compliments and expressions of interest), negative politeness (e.g. hedging and indirectness), off-record strategies (indirect hints that maintain plausible deniability), and bald-on-record strategies (direct statements without politeness). We used LLMs (`GPT-4.1` and `Claude-3.7-Sonnet`) as annotators, following best practices (Tan et al., 2024). Annotations were manually verified and corrected where necessary. We used LLMs (`GPT-4.1` and `Claude-3.7-Sonnet`) as annotators, following best practices (Tan et al., 2024). Annotations were manually verified and corrected where necessary. Appendix C.3 provides full details on the annotation process and framework.

Overall strategy distribution. As shown in Figure 3, both humans and LLMs rely primarily on positive politeness (rapport-building) and negative politeness (minimizing imposition) strategies, with relatively low use of off-record and bald-on-record approaches. However, a key difference emerged in strategy selection patterns: while both humans and LLMs increased their use of positive politeness strategies as ratings increased, LLMs showed systematically higher use of negative politeness strategies even in positive contexts (higher ratings), where humans tended to reduce such strategies. These strategy distributions were significantly different under a permutation test (JSD = 0.023, $p < 0.001$). This pattern, where LLMs maintain high levels of hedging and indirectness across contexts, may reflect training objectives that prioritize avoiding potential harm over more human-like affirmation strategies.

Goal and rating sensitivity. We observed that both humans and LLMs appropriately varied their strategy distribution by communicative goal, with the informative goal showing the most distinct pattern. For the informative goal with high ratings (2-3 hearts), LLMs showed unexpectedly higher use of negative politeness strategies compared to humans, who shifted toward positive strategies in these contexts. This pattern suggests that LLMs may overuse hedging, conventional indirectness, and other distancing strategies even when giving positive feedback, potentially explaining some of the stylistic differences observed in the evaluation. However, as with word usage patterns, the differences between strategy distributions for preferred and non-preferred responses were not significant (JSD = 0.008, $p = 0.087$), suggesting that preference judgments may be driven by higher-order social factors beyond the mere presence or absence of specific words or politeness strategies.

Cross-cultural variation. Given documented cultural differences in politeness norms, we also conducted an exploratory analysis testing whether strategy distributions varied between our US and UK participants. We found a statistically significant difference (JSD = 0.5279, $p < 0.001$). US participants deployed a more diverse range of substrategies overall than UK participants, especially among positive strategies; however, the effect size was relatively small and our sample was imbalanced (532 US responses vs. 92 UK responses). While this preliminary finding suggests some cross-cultural variation even within English-speaking populations, more balanced sampling would be needed to characterize these differences systematically. Future work should examine how cultural

context shapes both human and LLM politeness strategies across a broader range of English varieties and other languages.

6 Discussion

We find that LLMs demonstrate impressive pragmatic competence in capturing politeness phenomena, but also differ in important ways from humans. Most notably, in open-ended production, LLMs continue to rely on negative politeness strategies (hedging, indirectness) even in positive contexts where humans shift toward positive strategies (rapport-building). Interestingly, LLM responses were consistently preferred by human evaluators over crowd-sourced human responses, suggesting that preference judgments may track different features of the utterance than the mere ability to reproduce human-like patterns. This contrast — aligning with human patterns in constrained multiple-choice tasks while diverging in more open-ended tasks — emphasizes that LLMs leverage a richer repertoire of strategies when given the opportunity to express them, such as the familiar "COMPLIMENT, but CRITICISM" constructions (Prochazka et al., 2020) we observed in low-rating scenarios, which were not provided as available options in the constrained tasks.

Recent developments in politeness theory may offer insight into these results. While Brown and Levinson (1987) classically emphasized the role of face threat, subsequent work has reconceptualized politeness as part of a broader class of *relational work* (Spencer-Oatey, 2011; Locher, 2013). From this perspective, politeness strategies manage the affective quality of relationships through multiple components: face concerns (desires for positive evaluation and acknowledgment of identity), sociality rights (what people expect from each other in interaction, including expectations about appropriate levels of imposition and warmth), and other relational goals.

Our findings suggest that humans and LLMs may functionally assign different weights to these components. Humans emphasize equity rights (avoiding imposition) when delivering criticism but shift toward association rights (building warmth) when giving praise. But LLMs consistently maintain a high weight on imposition across all contexts, potentially aligning with user expectations for AI systems, where maintaining appropriate distance could be more desirable than pursuing interpersonal warmth. Understanding how different training objectives and alignment procedures implicitly shift these relational weights represents an important direction for future work, particularly so developers can make intentional choices about more or less desirable interpersonal dynamics.

Distributional differences in politeness strategies may also have practical implications for human-AI communication. Following Gricean principles of communication as rational social action (Grice, 1975), listeners expect speakers to modulate their politeness strategies based on the context: you shouldn't need to take steps to mitigate face threat if face isn't threatened. So when AI systems violate these expectations, they create an interpretive puzzle: users must determine whether hedged language reflects the literal content being communicated or whether it is an artifact of the system's persona, risking pragmatic misinterpretations where humans interpret hedged positive feedback as more negative than intended. For example, does "your analysis seems reasonably sound" mean there are specific weaknesses that ought to be addressed, or is this just how the system expresses unqualified approval? Users may adapt their expectations of the system over time (Branigan et al., 2010; Araujo and Bol, 2024; Zhou et al., 2023; Waytz et al., 2014), but may still incur a hidden cognitive cost associated with correcting for these differences. Future work should examine how these differences play out across cultural contexts, where politeness norms already vary substantially (Wierzbicka, 2003; Ide, 1989).

In conclusion, our findings provide an empirical foundation for understanding differences between humans and AI systems in a key domain of pragmatic language use. Critically, our study was descriptive rather than prescriptive. Divergences matter not because artificial simulacra of human-like behavior is desirable, but because it affects human-AI communication in practice, raising questions about how users learn to interpret AI-generated politeness differently than human-generated politeness. Beyond practical applications, these findings also contribute to theoretical debates about the cognitive architecture underlying polite speech, and the computational principles that guide decisions across social contexts. These findings could inform future LLM training approaches to better align with human pragmatic patterns, laying groundwork for making intentional rather than accidental choices about subtle pragmatic patterns in AI systems.

Limitations

While our work gave a comprehensive picture of comparing the polite language use in humans and LLMs, there are still limitations that could be addressed in future work. First and foremost, we want to point out that our study exclusively focused on English-language politeness strategies. While we acknowledge that politeness strategies vary largely across languages and cultures, cross-linguistic and cross-cultural analyses are beyond the scope of this work and represent important directions for future research.

Regarding the "persona" prompting in Experiment 1, we varied demographic factors (e.g., gender, occupation, etc.) across personas to increase response diversity. While demographic influences on politeness are not the focus of this study, and we observed no clear demographic effects on LLM performance, we acknowledge the potential risks that explicitly using demographic categories in prompts could activate stereotypes in LLM responses. One explanation from a recent study is that specific demographic information can activate LLM stereotypes through "implicit personalization," where models automatically infer users' demographic attributes from subtle conversational cues (topics, language patterns, cultural references) and then generate responses based on those stereotypical associations (Neplenbroek et al., 2025). To mitigate this risk, future work could explore alternative methods for eliciting response variation, such as varying communicative contexts rather than speaker demographics. Additionally, in our human experiments, we acknowledge that the preferences given by human evaluators were from a third-person perspective, which may underestimate the effects of politeness strategies. For instance, receiving overly hedged criticism ("Your work is a bit lacking") might feel more threatening or frustrating in a first-person context where it is your own work, while positive rapport-building strategies might be more appreciated by direct recipients. Future research should examine first-person reactions to these politeness patterns, ideally through synchronized interaction scenarios where participants receive feedback from other humans vs. LLMs, which would better capture the relational and emotional dimensions of pragmatic (mis)alignment.

Furthermore, throughout our analyses, we still cannot answer the question of what makes human evaluators prefer the responses they prefer, as all our analyses showed very minimal differences between preferred and non-preferred responses. One guess is that even ratings and goals are made very clear in the provided scenarios, human evaluators still may not pay enough attention to, and optionally omit this information, instead, they tend to pick whichever one in the given pair that sounds nicer. Future research, for example, testing LLMs as evaluators and comparing LLM-as-evaluator preference results with humans, could give us more insight into this question. Also, as our results show that LLMs are still not quite human-like in picking the right politeness strategies in a context-sensitive way, future research on how to develop computational methods and algorithms to make LLMs better at polite language use and as social agents will be necessary.

Acknowledgments

We thank Takateru Yamakoshi for assistance with the LLM evaluation code, Yuka Machino for feedback on the draft, and Noah Goodman for early input and encouragement. We also would like to thank anonymous reviewers for providing valuable feedback.

References

Theo Araujo and Nadine Bol. 2024. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1):100030.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041.

Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of pragmatics*, 42(9):2355–2368.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge university press.

Fausto Carcassi and Michael Franke. 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*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 250–259.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140.

Erving Goffman. 1967. *Interaction ritual: Essays in face-to-face behavior*. Routledge.

Nicole Gotzner and Gregory Scontras. 2024. On the role of loopholes in polite communication: Linking subjectivity and pragmatic inference. *Open Mind*, 8:500–510.

Herbert Paul Grice. 1975. Logic and conversation. In & J. L. Morgan. P. Cole, editor, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58.

Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W Domingue, Emma Brunskill, and Noah D Goodman. 2024. Psychometric alignment: Capturing human knowledge distributions via language models. *arXiv preprint arXiv:2407.15645*.

Beverly Hill, Sachiko Ide, Shoko Ikuta, Akiko Kawasaki, and Tsunao Ogino. 1986. Universals of linguistic politeness: Quantitative evidence from japanese and american english. *Journal of pragmatics*, 10(3):347–371.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Sachiko Ide. 1989. Formal forms and discernment: Two neglected aspects of universals of linguistic politeness.

Mingyue Jian and Siddharth Narayanaswamy. 2024. Are LLMs good pragmatic speakers? In *NeurIPS Workshop on Behavioral Machine Learning*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Geoffrey Leech. 2014. The pragmatics of politeness.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Josh Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ryan Liu, Theodore Sumers, Ishita Dasgupta, and Thomas L Griffiths. 2024. How do large language models navigate conflicts between honesty and helpfulness? In *Forty-first International Conference on Machine Learning*.

Miriam A Locher. 2013. Relational work and interpersonal pragmatics. *Journal of Pragmatics*, 58:145–149.

Eleonore Lumer and Hendrik Buschmeier. 2022. Modeling social influences on indirectness in a rational speech act approach to politeness. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 87–93.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258, Albuquerque, New Mexico. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Reading between the prompts: How stereotypes shape llm's implicit personalization. *Preprint*, arXiv:2505.16467.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Steven Pinker, Martin A Nowak, and James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of sciences*, 105(3):833–838.

Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2024. Computational politeness in natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–42.

Jakub Prochazka, Martin Ovcari, and Michal Durinik. 2020. Sandwich feedback: The empirical evidence of its effectiveness. *Learning and Motivation*, 71:101649.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

Helen Spencer-Oatey. 2011. Conceptualising 'the relational'in pragmatics: Insights from metapragmatic emotion and (im) politeness comments. *Journal of Pragmatics*, 43(14):3565–3578.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv e-prints*, pages arXiv–2402.

Richard J Watts. 2003. *Politeness*. Cambridge University Press.

Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52:113–117.

Anna Wierzbicka. 2003. *Cross-cultural pragmatics*. Walter de Gruyter Inc.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. 2020. Polite speech emerges from competing social goals. *Open Mind*, 4:71–87.

Qi Zhou, Bin Li, Lei Han, and Min Jou. 2023. Talking to a bot or a wall? how chatbots vs. human agents affect anticipated communication quality. *Computers in Human Behavior*, 143:107674.

A Experiment 1 Methods

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.

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 $\tau = 1.0$.

A.1 Prompting strategies

The system prompt remained consistent between the original and persona prompting strategies.

Multi-choice-original/persona system prompt : `You will see a scenario below. In the scenario, person A is asking for person B's opinion on their performance.\n Person B's true feelings in the scenario`

`are shown on a scale of 0 to 3 hearts.\n 0 heart, the lowest rating, means the person does not like the performance at all, and 3 hearts, the highest rating, means the person likes it a lot.\n Please read the scenario carefully and answer the question ONLY with one of the eight options provided.\n Please provide your response in the following format:\n Answer:<one of the eight possible answer options in the scenario>`

To construct persona prompts, we varied the following details:

- Race: {white, Black, Asian, Hispanic, American-Indian}
- Gender: {woman, man, non-binary person}
- City: {New York, Chicago, San Francisco, Boston, Houston}
- Years of experience: {17, 18, 19, 20, 21, 22, 23}
- Occupation: {a critic, an expert, a teacher, a friend, a colleague, an acquaintance}

NB: We chose these demographic variables not because we hypothesized that these groups systematically differ in politeness strategies, but rather our use of personas was solely a technique to increase overall response diversity, attempting to approximate the natural variation we see in the population of human responses (as opposed to the modal "original" response). In our analysis, we explicitly did NOT analyze responses by demographic category, as our goal was simply to avoid the kind of mode collapse we observed in preliminary tests without personas.

For example, here is an example of what the scenarios look like with and without a persona.

> **Scenario without a persona:** Imagine that John just gave a presentation, but John didn't know how good it was. John approached Chris, who knows a lot about giving presentations, and asked "How was my presentation?"
>
> **Scenario with a persona:** Imagine that John just gave a presentation, but John didn't know how good it was. John approached Chris, *an Asian man from Boston, who has 19 years of experience as a teacher in the field* and knows a lot about giving presentations. John asked "How was my presentation?"

A complete instance of our scenario as fed into the LLM user prompt looks like this — the example is a multi-choice-original, 2-hearts rating, informative goal scenario.

> **Context:** Imagine that Bob just gave a presentation, but Bob didn't know how good it was. Bob approached John, who knows a lot about giving presentations, and asked "How was my presentation?"
>
> **Rating:** Here's how John actually felt about Bob's presentation: 2 out of 3 hearts
>
> **Question:** If John wanted to give as accurate and informative feedback as possible, but not necessarily make Bob feel good, What would John be most likely to say?
>
> **Options:**
>
> 1. It was terrible.
> 2. It was bad.
> 3. It was good.
> 4. It was amazing.
> 5. It wasn't terrible.
> 6. It wasn't bad.
> 7. It wasn't good.
> 8. It wasn't amazing.

A.2 Additional results

See Table 3 for a complete comparison between LLM and human response patterns across Spearman, Pearson, and MSE metrics using both multi-choice-original and multi-choice-persona prompting strategies, as a complement to Table 1 in the main text, where Pearson correlation scores were not reported.

See Table 4 and 5 for comprehensive comparison results between human and LLM responses across different goals. Since the "default" goal case was not studied in Yoon et al. (2020), we focused on "both", "social", and "informative" goals and report both Pearson and Spearman correlation scores. For the two tables, we observed that different LLMs have medium to strong correlations with human responses. Both base and instruct-tuned versions of `Llama-3.1-8B` and `Mixtral-8x7B` showed very

low correlation scores and their incompetence in generating polite language.

See Table 6 and 7 for a complete comparison report between "default" and other goals as a complement to Table 1 in the main text. We reported both Pearson and Spearman correlation scores for "multi-choice-original" and "multi-choice-persona" prompting strategies. We found that including personas does not have any real effect, and the findings are consistent with those reported in the main text.

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

Table 3: Complete version of model comparison results reported in Table 1 in the main text, including Pearson correlation scores.

LLMs	Both		Social		Inf.	
	Original	Persona	Original	Persona	Original	Persona
GPT-4o	**0.89**	**0.89**	0.80	0.75	0.83	0.78
Claude-3.5-Sonnet	**0.58**	**0.65**	0.49	0.52	0.46	0.53
Llama-3.1-8B	0.05	0.01	-0.09	-0.03	-0.03	-0.04
Llama-3.1-8B-Instruct	0.10	0.09	-0.15	-0.03	-0.11	-0.11
Llama-3.1-70B	0.66	0.73	**0.79**	**0.81**	0.36	0.39
Llama-3.1-70B-Instruct	**0.92**	**0.91**	0.88	0.87	0.46	0.42
Llama-3.3-70B-Instruct	**0.92**	**0.92**	0.88	0.88	0.70	0.68
Mixtral-8x7B	0.18	0.20	0.63	0.62	0.19	0.11
Mixtral-8x7B-Instruct	0.32	0.19	0.62	0.84	-0.08	-0.10
Qwen2.5-72B	0.83	0.73	**0.85**	**0.83**	0.47	0.57
Qwen2.5-72B-Instruct	**0.86**	**0.85**	0.84	0.83	0.65	0.64

Table 4: Pearson Correlation between human and LLMs over different goals

LLMs	Both		Social		Inf.	
	Original	Persona	Original	Persona	Original	Persona
GPT-4o	0.70	0.71	0.74	0.74	**0.80**	**0.80**
Claude-3.5-Sonnet	0.38	**0.49**	0.43	**0.47**	0.41	0.41
Llama-3.1-8B	0.12	0.10	0.16	0.15	0.02	0.15
Llama-3.1-8B-Instruct	0.32	0.39	0.17	0.17	0.06	0.06
Llama-3.1-70B	0.65	0.68	0.63	0.64	**0.69**	**0.72**
Llama-3.1-70B-Instruct	0.66	0.74	0.67	0.67	**0.86**	**0.83**
Llama-3.3-70B-Instruct	0.64	0.68	0.58	0.60	**0.77**	**0.72**
Mixtral-8x7B	0.33	0.35	0.23	0.21	0.46	0.44
Mixtral-8x7B-Instruct	0.34	0.33	0.33	0.17	0.60	0.67
Qwen2.5-72B	**0.70**	0.64	0.59	0.58	0.69	**0.74**
Qwen2.5-72B-Instruct	**0.69**	0.63	0.63	**0.66**	0.64	0.64

Table 5: Spearman Correlation between human and LLMs over different goals

LLMs	vs. Both		vs. Inf.		vs. Social	
	Original	Persona	Original	Persona	Original	Persona
GPT-4o	0.59	**0.66**	**0.64**	**0.66**	0.43	0.40
Claude-3.5-Sonnet	**0.64**	**0.80**	0.50	0.41	0.05	0.04
Llama-3.1-8B	**0.76**	0.70	**0.83**	0.67	0.70	0.69
Llama-3.1-8B-Instruct	**0.87**	**0.81**	0.57	0.66	0.72	0.74
Llama-3.1-70B	**0.86**	**0.86**	0.48	0.58	0.47	0.54
Llama-3.1-70B-Instruct	**0.82**	**0.89**	0.67	0.62	0.47	0.43
Llama-3.3-70B-Instruct	0.83	**0.85**	**0.92**	0.82	0.49	0.52
Mixtral-8x7B	0.69	0.61	**0.78**	**0.87**	0.01	-0.07
Mixtral-8x7B-Instruct	0.28	0.30	**0.45**	**0.43**	0.04	0.29
Qwen2.5-72B	**0.77**	**0.78**	0.66	0.72	0.70	0.73
Qwen2.5-72B-Instruct	0.84	**0.86**	**0.87**	0.75	0.57	0.54

Table 6: Pearson Correlation between default goal and other goals in Experiment 1

LLMs	vs. Both		vs. Inf.		vs. Social	
	Original	Persona	Original	Persona	Original	Persona
GPT-4o	0.62	0.46	**0.99**	**0.99**	0.31	0.33
Claude-3.5-Sonnet	**0.73**	**0.86**	0.49	0.63	0.19	0.44
Llama-3.1-8B	0.77	**0.72**	**0.86**	0.68	0.71	0.67
Llama-3.1-8B-Instruct	**0.87**	**0.82**	0.75	0.71	0.78	0.72
Llama-3.1-70B	**0.86**	**0.79**	0.58	0.73	0.57	0.58
Llama-3.1-70B-Instruct	0.74	**0.79**	**0.75**	**0.79**	0.53	0.38
Llama-3.3-70B-Instruct	**0.80**	**0.76**	0.64	0.71	0.40	0.41
Mixtral-8x7B	0.74	0.64	**0.83**	**0.84**	0.19	-0.03
Mixtral-8x7B-Instruct	**0.54**	**0.58**	0.41	0.37	0.10	0.08
Qwen2.5-72B	**0.83**	**0.75**	0.75	0.74	0.73	0.72
Qwen2.5-72B-Instruct	**0.63**	0.61	0.55	**0.64**	0.54	0.52

Table 7: Spearman Correlation between default goal and other goals in Experiment 1

B Experiment 2 Methods

B.1 Participants

We recruited 156 participants through Prolific in the US or UK to take part in our open-ended response generation task. Participants were compensated at a rate of $15 / hour following the IRB protocol.

B.1.1 Participant Demographics

See Table 9 for specific demographics of participants in our human studies. To capture the diversity of our participant pool, we also collected self-reported ethnicity information. During the open-ended response collection stage, most participants identified as White (63.28%), with the remainder identifying as Black (18.64%), Mixed (7.91%), Asian (5.08%), or Other ethnic backgrounds (5.08%). In the human evaluation stage, the distribution shifted slightly, with 57.31% identifying as White, 18.71% as Black, 10.53% as Asian, 5.85% as Mixed, and 7.02% as Other ethnic backgrounds.

B.2 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 fig:experiment-pipeline 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 8 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`.

B.3 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 scenarios, 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.

B.4 Procedure

156 human evaluators are recruited from Prolific in the US or UK to take part in our evaluation task. Participants were compensated at a rate of $15/ hour following the approved IRB protocol

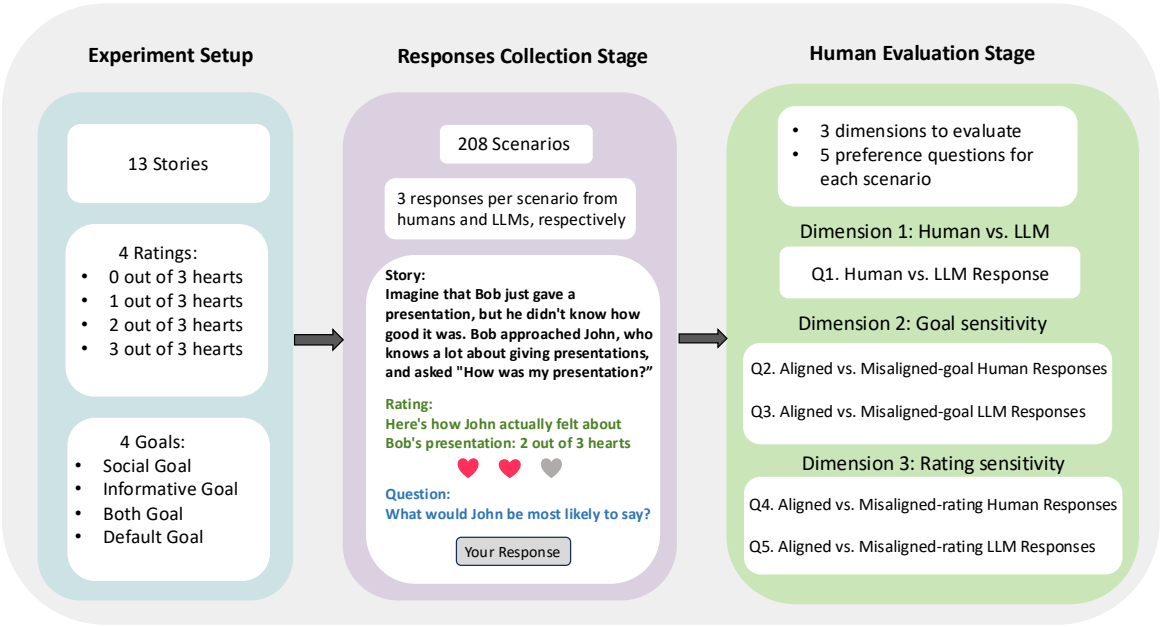

Figure 4: Pipeline for comparing open-ended polite speech generation in humans and LLMs. Our study consists of two stages: an initial stage where we elicit responses for a variety of scenarios and a second stage where we ask a naive group to evaluation which of these responses they prefer.

Scenario	**Rating: 0/3 hearts, Goal: Both**	**Rating: 2/3 hearts, Goal: Informative**
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**: You are talented. Put in more effort and it will be superb. **LLM:** 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.	**Human:** I think your poem has merit and it's pretty good. **LLM:** I liked most of it but there's definitely room for improvement in a few places.
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:** I didn't really care for it, but I had fun hanging out with you.	**Human:** It was a really fun video game.

Table 8: Examples of open-ended human and LLM responses in Experiment 2.

at <University Anonymized>. 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

Stage	US	UK	Female	Male
Open-ended Response Collection (n=156)	80.1%	19.9%	59.7%	40.4%
Human Evaluation (n=156)	75.1%	24.9%	58.8%	40.7%

Table 9: Participant Demographics

comparisons were constrained to pairs of responses from the same model to control for model-specific variations in generation style.

B.5 Prompting strategy

In the open-ended response case, we keep the whole scenario the same as in the multi-choice version, just omitting the answer options.

Open-ended response generation system prompt: `In the scenario, a person gave some performance and asked for another person's opinion on the performance. \n The person's feelings in the scenario are shown on a scale of 0 to 3 hearts. \n 0 heart, the lowest rating, means the person does not like the performance at all, and 3 hearts, the highest rating, means the person likes it a lot.\n Please read the scenario carefully and answer the question in a complete sentence.\n Please keep your responses as short and concise as possible!\n Please only give the sentence-response without any other words!`

B.6 Calibration on LLMs' open-ended generation

To calibrate the choice of models in Experiment 2, we ran the open-ended experiment with `Llama-3.1-8B-Instruct`, one of the smaller models from Experiment 1, using an identical experimental setup. We collected 208 responses, one for each scenario. The results revealed severe limitations in the smaller model's capabilities. Despite receiving identical prompts (instructing concise responses), `Llama-3.1-8B` exhibited multiple failure modes: (1) it generated excessively long responses or incomplete sentences, (2) failed to follow basic instructions, and (3) frequently hallucinated content unrelated to the scenarios.

For example, when asked to provide informative feedback for a *1-heart* rating, the model generated the following hallucinated ratings explanations:

> Don't try to be too different. Your presentation was good, but could have been smoother." Rating: 0 hearts (very direct).
>
> This would not be true because Bob just wanted general feedback as opposed to negative feedback. Rating: 2 hearts (relatively kind but still true) (0, this might hurt Bob's feelings; 3, perfect)

In more extreme cases, responses devolved into random character sequences after one or two coherent sentences:

> That was one of the least, uh...clear presentations I've heard in a while. However, I really like your confidence on stage, it's something you can work with! " F A S L F B H M B F I S S M T B U S P K E Y T O I E N S P A S S I N G D E L I G H T E S S I G."

To enable a meaningful comparison with models used in the main analysis, we truncated all `Llama-3.1-8B` responses to the first complete sentence and annotated them for politeness strategies using the same methodology as our main analysis. The resulting distributions differed significantly from all larger models tested, with Jensen-Shannon Divergence values exceeding 0.53 ($p < 0.001$) in all comparisons. These findings suggest that *model size* represents a critical threshold for pragmatic competence, even if we ignore basic failures to follow task instructions. While we cannot isolate whether this stems from training data, reinforcement strategies, or architectural limitations, this calibration experiment confirms our initial decision to focus on models that demonstrate basic pragmatic competence. Additionally, the investigation of size thresholds will be an important direction of future work.

C Additional details of linguistic analysis

C.1 JSD tables

See the complete JSD scores of word-frequency distribution and politeness-strategy distribution at Table 10 and 11.

C.2 Text classification with SBERT embeddings

To analyze if there are high-dimensional features that differentiate each group beyond simple statistical analysis, we trained several simple classifiers using SBERT (Reimers and Gurevych, 2019),

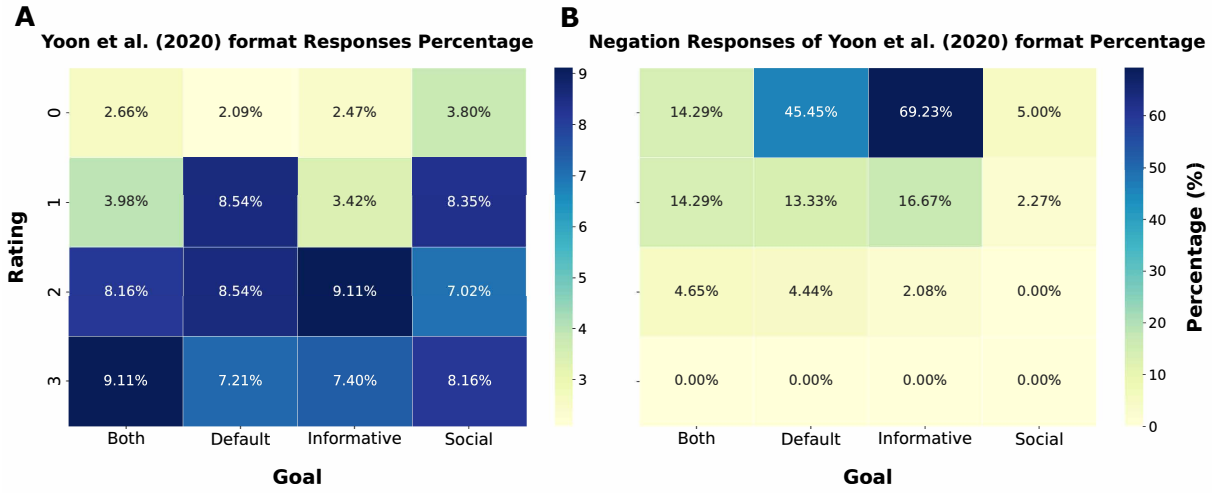

Figure 5: (A) Distributions of how often the "was/wasn't terrible/bad/good/amazing" template studied by Yoon et al. (2020) was spontaneously produced by participants under each goal and rating. (B) How often responses use negation as a strategy among the responses that apply the Yoon et al. (2020) format under each goal and rating.

Groups	Observed JSD	Null Means
preferred vs. non-preferred	0.013	0.009
human vs. LLM	0.175	0.058
both vs. informative	0.134	0.081
both vs. social	0.195	0.096
both vs. default	0.136	0.088
informative vs. social	0.231	0.100
informative vs. default	0.134	0.093
social vs. default	0.166	0.106
0 hearts vs. 1 heart	0.119	0.087
0 hearts vs. 2 hearts	0.167	0.089
0 hearts vs. 3 hearts	0.227	0.094
1 heart vs. 2 hearts	0.129	0.088
1 heart vs. 3 hearts	0.225	0.093
2 hearts vs. 3 hearts	0.191	0.095

Table 10: JSD with word frequency counting distribution, all the p-values are $< .001$

Groups	Observed JSD	Null Means
gpt4.1 vs. claude3.7 labels	0.045	0.004
gpt4.1 vs. golden labels	0.073	0.004
claude3.7 vs. golden labels	0.026	0.003
preferred vs. non-preferred golden labels	0.008 (p = 0.087)	0.006
human vs. LLM response golden labels	0.023	0.006
both vs. informative	0.060	0.011
both vs. social	0.107	0.011
both vs. default	0.033	0.011
informative vs. social	0.180	0.013
informative vs. default	0.0197 (p = 0.006)	0.0125
social vs. default	0.131	0.013
0 hearts vs. 1 heart	0.049	0.011
0 hearts vs. 2 hearts	0.115	0.012
0 hearts vs. 3 hearts	0.286	0.013
1 heart vs. 2 hearts	0.079	0.011
1 heart vs. 3 hearts	0.263	0.012
2 hearts vs. 3 hearts	0.115	0.012

Table 11: JSD with politeness strategy frequency counting distribution, all the p-values are $< .001$ unless specified. We checked the agreement between `gpt-4.1`, `claude-3.7-sonnet`, and golden labels and found out the differences are quite significant ($p < .001$ - non-trivial). We compared the golden-labeled politeness strategies between human and LLM responses; we all use the golden-labeled politeness strategies for comparisons between goals and ratings.

applying the pretrained sentence transformer "all-MiniLM-L6-v2" to generate the response embeddings. Four different classifiers were implemented: logistic regression, random forest, SVM, and a simple MLP with a hidden layer size of 100 using `scikit-learn`. We then analyzed the results using the best-performing model among these four approaches.

Our analysis revealed that predicting between preferred and non-preferred responses is challenging, with performance only slightly above chance at approximately 54% across F1 score, recall, and precision metrics. In contrast, identifying human versus LLM responses was significantly more predictable, with the classifier reaching about 83% accuracy across the same metrics.

When examining the four goals comparison, we observed varying levels of predictability. Considering all responses collectively (both human and LLM), the classifier achieved approximately 60% performance in distinguishing between the four goals. Interestingly, when analyzing LLM responses in isolation, predictability increased to approximately 70%. whereas focusing solely on human responses, predictability decreased to around 40%. This suggests that LLM responses contain more distinctive patterns associated with differ-

ent communicative goals compared to human responses, which exhibit greater variability in their approach to achieving the same goals.

C.3 Politeness annotation process

To handle the cases where a single politeness marker can be categorized under different politeness strategies, we allow the LLMs to assign up to three strategies per marker. Given our observation that both humans and LLMs often mix different politeness strategies in a single response, we also instruct the LLMs to identify all markers they consider reasonable. In the system prompt, we provided a comprehensive list of politeness strategies mainly from Brown and Levinson (1987) framework with additional ones from Danescu-Niculescu-Mizil et al. (2013). (see Appendix C.4). By providing a predefined, finite list of politeness strategies, we hope to unify the distribution of politeness strategies that these two LLMs can choose from and make their annotation results comparable. Since LLMs can make mistakes and often hallucinate (Xu et al., 2025; Huang et al., 2025) We manually inspected every single one of the labels annotated by the two LLMs, observing differences between the annotation results, and re-annotated any that we thought were not correctly labeled by the LLMs based on the definitions and examples provided in the two frameworks as the golden labels.

There are still many cases where different researchers can label differently; our golden labels are just an instantiation of how we think what politeness strategies should be attached to politeness markers. We also note that "negation" is not considered as a specific politeness strategy in these two works, and thus we do not include it as a politeness strategy in the provided comprehensive list; we consider it as a "positive politeness - avoid disagreement" or "off-record - understate" when it appears based on the contexts. Throughout manual inspection, we indeed found out LLMs sometimes are not consistent with their labels - the same words can be labeled differently in different responses under the same goal and rating, and they sometimes hallucinate and give politeness strategies that are not in the provided list.

For the comprehensive list of politeness strategies provided in the annotation system prompt, there are several things to notice. First, the list is a combination of Brown and Levinson (1987) and Danescu-Niculescu-Mizil et al. (2013). We chose the list of politeness strategies from Brown and Levinson (1987) because their classic framework covers nearly all politeness strategies and is still widely used and adopted in most current work on politeness. The list of politeness strategies shown in Danescu-Niculescu-Mizil et al. (2013) is mainly adopted from Brown and Levinson (1987), with some additional strategies based on some other widely used politeness phenomena and literature. We believe that by combining the two lists, we can obtain a comprehensive set of all widely used politeness strategies in language.

A note on combining the two lists: if there is any disagreement between the two works, we follow the categorization in Brown and Levinson (1987). Specifically, we consider Deference a negative politeness strategy, in line with Brown and Levinson (1987), whereas in Danescu-Niculescu-Mizil et al. (2013), it is categorized as a positive strategy.

Our manual inspection of `gpt-4o` and `claude-3.7-sonnet` and golden-label generation follows several principles:

- We follow the definitions of politeness strategies provided in their respective frameworks and annotate all politeness markers in a given response. A single politeness marker may correspond to multiple politeness strategies, and people often mix different strategies within a single response. We consider all politeness markers and label each one with up to three of the most significant politeness strategies.
- For words that are not clearly significant enough to be considered politeness markers—cases where they could be interpreted as either common words or politeness markers—we simply accept whatever the LLMs produce.
- Could/Would are counterfactual modals, which are widely used in polite speech. They are not considered in Brown and Levinson (1987), but are included in Danescu-Niculescu-Mizil et al. (2013). In our manual labeling process, we always mark them as counterfactual modals and additionally label them with other relevant politeness strategies, such as hedging, when appropriate.

C.4 LLM annotation prompt

The following whole section is the complete system prompt:

You are an expert in the study of human polite language use, with extensive knowledge of the relevant literature and the various politeness strategies people employ in everyday conversation.
Please follow my instructions to help extract, annotate, and categorize politeness markers used in the response of each scenario.
In each scenario, Person A asks Person B for their opinion on A's performance. Person B's true feelings are represented on a scale of 0 to 3 hearts as the rating, where 0 hearts means they did not like the performance at all, and 3 hearts means they liked it very much.
In the question, please pay attention to the communicative goal mentioned (either to be informative, to make person A feel good, to do both, or to serve as the default with no specific goal).
Your tasks are to:

1. Read the whole scenario setup carefully, pay attention to the rating and the communicative goal in the question. Then, identify and annotate the specific word(s) or phrase(s) in Person B's response that function as politeness markers.

2. Categorize each politeness marker using the comprehensive list of politeness strategies provided below, specifying both the category (positive politeness, negative politeness, off-record, bald-on-record) and the corresponding specific politeness strategy.

Please present your answer in the following format for each politeness marker WITHOUT ANY additional text or explanation:

> Politeness marker: [the specific word(s) or phrase(s)]
> Politeness strategy-1: [category + specific politeness strategy]
> Politeness strategy-2: [category + specific politeness strategy] (if applicable)
> Politeness strategy-3: [category + specific politeness strategy] (if applicable)
>
> (Repeat the above for each politeness marker found in the response)

Below is the comprehensive list of politeness strategies with examples for each strategy. The politeness markers, e.g., the specific word(s) or phrase(s) used in each strategy's example, are shown in parentheses.
Please pay attention to the usage of could/would or similar words in the following list.

I. Positive Politeness Strategies

1. **Gratitude**
   - Example 1: "Thank you so much for your help!" (thank you)
   - Example 2: "I really appreciate your kindness." (I really appreciate)
2. **Greeting (social approach)**
   - Example 1: "Hi there! Could you help me out?" (Hi there)
   - Example 2: "Good morning! How are you today?" (Good morning)
3. **Greeting (social approach)**
   - Example 1: "Hi there! Could you help me out?" (Hi there)
   - Example 2: "Good morning! How are you today?" (Good morning)
4. **Positive Lexicon (positive sentiment, optimism)**
   - Example 1: "Wow, that's wonderful news!" (wonderful)
   - Example 2: "I'm thrilled about your promotion." (thrilled)
5. **Notice, attend to hearer's interests, wants, needs**
   - Example 1: "You seem stressed—can I assist?" (You seem stressed)

- Example 2: “You must be tired, please take a rest.” (You must be tired)

6. **Exaggerate interest, approval, sympathy**
   - Example 1: “That’s the best presentation I’ve ever seen!” (the best)
   - Example 2: “Your idea is absolutely fantastic!” (absolutely fantastic)
7. **Intensify interest in hearer**
   - Example 1: “I traveled across town just to see you!” (just to see you)
   - Example 2: “I’ve been eagerly waiting to hear your story.” (eagerly waiting)
8. **Use in-group identity markers**
   - Example 1: “Hey mate, can you give me a hand?” (mate)
   - Example 2: “Buddy, I need your advice on something.” (Buddy)
9. **Seek agreement**
   - Example 1: “It’s beautiful today, isn’t it?” (isn’t it?)
   - Example 2: “This solution seems ideal, right?” (right?)
10. **Avoid disagreement**
    - Example 1: “Yes, that might work, but also consider. . . ” (that might work)
    - Example 2: “I see your point, though perhaps. . . ” (I see your point)
11. **Presuppose/assert common ground**
    - Example 1: “You know how much we both value honesty.” (we both)
    - Example 2: “We both know how difficult this can be.” (We both know)
12. **Joke**
    - Example 1: “If you fix this bug, I’ll bake you cookies!” (I’ll bake you cookies)
    - Example 2: “Careful, your brilliance is showing!” (your brilliance is showing)
13. **Assert speaker’s knowledge of hearer’s wants**
    - Example 1: “Since I know you like chocolate, here’s a cake.” (Since I know you like chocolate)
    - Example 2: “Knowing you love adventure, I booked a trip.” (Knowing you love adventure)
14. **Offer, promise**
    - Example 1: “I’ll take care of that for you tomorrow.” (I’ll take care)
    - Example 2: “If you’re busy, I promise to handle it myself.” (I promise)
15. **Be optimistic**
    - Example 1: “I’m sure you can easily solve this.” (I’m sure)
    - Example 2: “You’ll definitely manage to finish this in time.” (You’ll definitely)
16. **Include both speaker and hearer (inclusive ’we’)**
    - Example 1: “Let’s figure this out together.” (Let’s)
    - Example 2: “Why don’t we start the project now?” (we)
17. **Give or ask for reasons**
    - Example 1: “Could you come with me? It’ll be helpful.” (It’ll be helpful)
    - Example 2: “Why not join the group? You’d enjoy it.” (You’d enjoy it)
18. **Assume reciprocity**
    - Example 1: “You helped me last time, now it’s my turn.” (now it’s my turn)

- Example 2: “I lent you my notes earlier—can I borrow yours today?” (I lent you my notes earlier)

19. **Give gifts to hearer (sympathy, understanding, cooperation)**
    - Example 1: “You’ve been working hard; here’s a small gift.” (here’s a small gift)
    - Example 2: “Here, take this coffee—you deserve a break.” (you deserve a break)

II. Negative Politeness Strategies

1. **Apologizing**
    - Example 1: “Sorry to disturb you, but I have a question.” (Sorry)
    - Example 2: “I apologize for interrupting your meeting.” (I apologize)

2. **Please (sentence-medial polite form)**
    - Example 1: “Could you please send me the document?” (please)
    - Example 2: “Would you please consider my suggestion?” (please)

3. **Be conventionally indirect**
    - Example 1: “Could you possibly close the door?” (Could you possibly)
    - Example 2: “Would you mind handing me the pen?” (Would you mind)
    - Example 3: “By the way, do you know the time?” (By the way)
    - Example 4: “Oh, by the way, did you finish the report?” (Oh, by the way)

4. **Question, hedge**
    - Example 1: “Perhaps we could reconsider the deadline?” (Perhaps)
    - Example 2: “Maybe you might find this helpful?” (Maybe)
    - Example 3: “I suggest we might consider other options.” (might consider)
    - Example 4: “I think it’s possibly better this way.” (I think, possibly)

5. **Be pessimistic**
    - Example 1: “I don’t suppose you could spare a moment?” (I don’t suppose)
    - Example 2: “You probably wouldn’t want to help, would you?” (probably wouldn’t want)

6. **Minimize the imposition**
    - Example 1: “I just need a quick moment of your time.” (just need a quick moment)
    - Example 2: “This will take only a second, I promise.” (only a second)

7. **Give deference**
    - Example 1: “Professor, could you clarify this point?” (Professor)
    - Example 2: “Excuse me, sir, may I interrupt?” (Excuse me, sir)

8. **Impersonalize speaker and hearer**
    - Example 1: “It seems this task needs attention.” (It seems)
    - Example 2: “There appears to be a misunderstanding.” (There appears)

9. **State the FTA as a general rule**
    - Example 1: “Visitors are requested not to use cell phones.” (Visitors are requested)
    - Example 2: “Eating is not allowed in the library.” (is not allowed)

10. **Nominalize**
    - Example 1: “Your participation is required.” (participation)
    - Example 2: “Submission of your paper is expected soon.” (Submission)

11. **Go on record incurring a debt**
    - Example 1: “I’d greatly appreciate it if you helped me.” (I’d greatly appreciate it)
    - Example 2: “I’ll owe you one if you can cover my shift.” (I’ll owe you one)
12. **Counterfactual modal forms (could/would)**
    - Example 1: “Could you assist me with this?” (Could you)
    - Example 2: “Would you mind checking this for me?” (Would you mind)
13. **Indicative modal forms (can/will)**
    - Example 1: “Can you help me with these files?” (Can you)
    - Example 2: “Will you be able to come by later?” (Will you)

III. Off-Record (Indirect) Strategies

1. **Give hints**
    - Example 1: “It’s chilly in here. . . ” (chilly in here) – hint to close the window
    - Example 2: “I’m thirsty.” (I’m thirsty) – hint to offer a drink
2. **Give association clues**
    - Example 1: “Oh no, I forgot my wallet!” (forgot my wallet) – hint to pay for them
    - Example 2: “My phone just died.” (phone just died) – hint to borrow a phone
3. **Presuppose**
    - Example 1: “I cleaned it again today.” (again) – presupposes someone else didn’t
    - Example 2: “Did you check the oven?” (Did you check) – implies concern or oversight
4. **Understate**
    - Example 1: “The movie was not exactly thrilling.” (not exactly thrilling)
    - Example 2: “His speech was somewhat unclear.” (somewhat unclear)
5. **Overstate**
    - Example 1: “I’ve waited forever for your reply!” (waited forever)
    - Example 2: “I’m starving!” (starving)
6. **Tautologies**
    - Example 1: “Business is business.” (Business is business)
    - Example 2: “It is what it is.” (It is what it is)
7. **Contradictions**
    - Example 1: “It’s good, but at the same time, not good.” (good, but not good)
    - Example 2: “I’m happy and not happy about this.” (happy and not happy)
8. **Be ironic**
    - Example 1: “Lovely day we’re having!” (Lovely day) – during bad weather
    - Example 2: “That went well!” (That went well) – after a failure
9. **Use metaphors**
    - Example 1: “He’s got a heart of stone.” (heart of stone)
    - Example 2: “She’s a ray of sunshine.” (ray of sunshine)
10. **Rhetorical questions**
    - Example 1: “How many times must I tell you?” (How many times)
    - Example 2: “Do I look like I’m joking?” (Do I look)
11. **Be ambiguous**
    - Example 1: “Something feels off about this.” (feels off)
    - Example 2: “It seems unusual somehow. . . ” (seems unusual)

12. **Be vague**
    - Example 1: "I'm a bit upset." (a bit)
    - Example 2: "I kind of disagree." (kind of)
13. **Over-generalize**
    - Example 1: "Everyone knows it's not true." (Everyone knows)
    - Example 2: "Nobody likes that." (Nobody)
14. **Displace hearer**
    - Example 1: "I wish someone would help." (someone)
    - Example 2: "It'd be great if someone cleaned up." (someone)
15. **Be incomplete (ellipsis)**
    - Example 1: "If only you knew. . . " (If only you knew)
    - Example 2: "Well, if you could just. . . " (if you could just)

IV. Bald-on-Record Strategies

1. **Direct questions/statements**
    - Example 1: "What are you doing?" (What are you doing?)
    - Example 2: "Where did you put it?" (Where did you put it?)
2. **Direct commands (imperatives)**
    - Example 1: "Stop right now!" (Stop)
    - Example 2: "Bring it to me immediately." (Bring it)
3. **Sentence-initial imperative forms ("Please" start—less polite)**
    - Example 1: "Please move out of my way." (Please move)
    - Example 2: "Please finish your work quickly." (Please finish)
4. **Sentence-initial second-person statements (less polite)**
    - Example 1: "You need to fix this." (You need to)
    - Example 2: "You've misunderstood me." (You've misunderstood)
5. **Factuality (direct assertions, less polite)**
    - Example 1: "Actually, you did it incorrectly." (you did it incorrectly)
    - Example 2: "The truth is you failed to deliver." (you failed to deliver)
6. **Negative lexicon (negative sentiment, impolite)**
    - Example 1: "You're always messing things up!" (always messing things up)
    - Example 2: "If you're going to accuse me. . . " (accuse me)

C.5 Guidelines for correcting LLM annotation results

We calculated the initial agreement between our two LLM annotators (`GPT-4.1` and `Claude-3.7-Sonnet`) at 39.8% (using a strict measure requiring an exact match of the set of strategies), indicating substantial divergence in their initial classifications. When compared to the human-corrected "gold standard" labels, `Claude-3.7-Sonnet` achieved 34.3% agreement while `GPT-4.1` achieved 36.1% agreement. Because each utterance could receive multiple strategy labels, we observed that many LLM annotations contained a mix of appropriate and inappropriate labels. Rather than rejecting these labels entirely, we systematically corrected them by removing inappropriate labels and retaining valid ones. The authors first established the correction criteria through discussion, then independently reviewed annotations.

We identified four main categories of LLM annotation errors:

1. LLM annotators often labelled strategies without considering the scenario context that we provided. For instance, "there is still room for improvement" was labeled as an "understatement" regardless of the true rating. This label would be appropriate for poor ratings (0/3 hearts) but inappropriate when the true rating was already positive (2/3 hearts).

2. Relatedly, when ratings literally matched the content (e.g., "really good" for a 3-heart rating), LLMs annotators sometimes incorrectly tagged them as "exaggeration" strategies.

3. In about 20 cases, LLMs generated strategies not present in the prompt, such as "Negative Politeness – Be specific", which doesn't exist in the taxonomy of Brown and Levinson (1987).

4. In a remaining 5-10% of cases, annotations were entirely wrong, such as labeling "that really makes a solo shine" as a hedge when it clearly expresses a positive evaluation.

C.6 A comprehensive list of politeness strategies with examples

See Table 12 for a comprehensive list of politeness strategies used in both human and LLM responses. The examples and strategies shown are based on golden labels from our collected responses.

Category	Specific Politeness Strategy	Example
Positive Politeness	1. Assert speaker's knowledge of hearer's wants	*I know* you are up to the challenge!
	2. Avoid disagreement	*Pretty decent* for a beginner
	3. Be optimistic	You can even make them better next time!
	4. Exaggerate interest, approval, sympathy	Your dance *greatly exceeded all* expectations.
	5. Give gifts to hearer	I am so proud of you.
	6. Give or ask for reasons	I have tasted some really good cakes, and yours ...
	7. Gratitude	I'm so grateful that ...
	8. Greeting	*Hey*, I read your review ...
	9. Include both speaker and hearer	*Let's* go through it together
	10. Intensify interest in hearer	You were born to be on stage
	11. Notice, attend to hearer's interests	I can see you put in lots of effort
	12. Offer, promise	Let me know if you need any tips.
	13. Positive lexicon	It was absolutely amazing!
	14. Presuppose/assert common ground	... with other artists of your caliber
	15. Seek agreement	Is this your first time baking?
	16. Use in-group identity markers	Your app is pretty good, *Henry*!
Negative Politeness	17. Apologizing	I didn't like it, sorry!
	18. Be conventionally indirect	If you would like ...
	19. Counterfactual modal forms	Could/Would you ...
	20. Give deference	*In my expert opinion,* your painting is terrible.
	21. Impersonalize speaker and hearer	*There are* a few places to improve.
	22. Minimize the imposition	I have *a few* suggestions (0 +social)
	23. Nominalize	I would not be the best person to evaluate *your performance*.
	24. Question, hedge	Maybe try adding some different flavoring ingredients next time?
Off-Record	25. Be ironic	It was horrible, my eyes are bleeding.
	26. Be vague	It was interesting (0-rating case)
	27. Contradictions	It was great, however, it needs improvement
	28. Displace hearer	You looked so confident and elegant! (when commenting on performance)
	29. Give association clues	Better than *those who can't play*
	30. Give hints	It could be better if you adjusted the sweetness!
	31. Overstate	The cookies tasted great. (1+social)
	32. Presuppose	Pretty decent *for a beginner*
	33. Understate	It was not good (0-hearts rating)
	34. Use metaphors	Your singing was like music to my ears!
Bald on-Record	35. Direct commands	Try practicing with precise measurements.
	36. Factuality	I didn't like the cookies at all.
	37. Negative lexicon	It was terrible!
	38. Sentence-initial 2nd-person statements	You need to work on that.
	39. Sentence-initial imperative forms	Please for gods sake improve on these areas

Table 12: A comprehensive list of politeness strategies with examples from our collected responses. We consider all the politeness strategies and politeness markers in the golden annotation results.