International Journal of English Language and Linguistics Research, 14, (2) 1-13, 2026
Print ISSN: 2053-6305(Print)
Online ISSN:2053- 6313(online
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

# Politeness Strategies in Human–AI Interaction: A Comparative Pragmatic Analysis of Arabic Responses Produced by ChatGPT and Human Participants

**Mafaz Hatem Ouda**
The General Directorate of Education in Misan Province,
Ministry of Education, Iraq

doi: https://doi.org/10.37745/ijellr.13/vol14n2113 Published September 28, 2026

**Citation**: Ouda M.H. (2026) Politeness Strategies in Human–AI Interaction: A Comparative Pragmatic Analysis of Arabic Responses Produced by ChatGPT and Human Participants, *International Journal of English Language and Linguistics Research*, 14, (2) 1-13

**Abstract**: *Even though conversational large language models are now often involved in socially relevant communications; it is still unclear whether their polite presentation is tuned to the context in a similar way as human language. The present study contrasts the pragmatic implementation of politeness in Arabic responses elicited by a set of 15 written discourse-completion tasks which differ in social distance, relative power, and degree of imposition, between human participants and ChatGPT. There were 660 valid human answers and 1 standardized ChatGPT answer per scenario in the human corpus, and one standardized ChatGPT answer per scenario in the exploratory AI corpus. Analysis is based on Brown's and Levinson's politeness theory and speech-act coding along with directness of request, supportive moves, modal realization, and length of response in comparison. Conventionally indirect forms were most frequent (53.7%) in 352 human requests, followed by direct requests (32.7%), hints (7.7%) and non-target realizations (6.0%); Cochran's $Q(7)=60.66$, $p<.001$. ChatGPT employed the conventionally indirect form in each of the 8 request situations and averaged out more supportive/mitigating devices per request (3.75 vs. 2.49). In all 15 scenarios the AI realization was found to be congruent with the modal human category in 9 cases (60.0%) and longer than the mean of the corresponding human scenario in all scenarios. The results show a pragmatic convergence, but systematic preference for elaboration, indirectness, and risk-averse mitigation in the responses. Human politeness was less rigid and contextually efficient. Note that the AI corpus consists of one generation per scenario, so the contrast with the human-generated one is an exploratory one and shouldn't be interpreted as a fixed distribution of the model's behaviour.*

**Keywords:** politeness, pragmatics, ChatGPT, large language models, human–AI interaction, Arabic, discourse completion test

## INTRODUCTION

Large language models (LLMs) have transformed the way people interact with computers, moving from a command-driven relationship to one that is based on ongoing conversation. Their products can

be grammatically correct, coherent and socially supportive and pragmatics (more than just being grammatically correct) are central to their evaluation. For instance, when providing advice, ChatGPT responses have been found to be highly balanced and empathetic, and are often more elaborate than human responses (Howe et al., 2023). Dialogue readiness systems also stress that it is not only grammar that matters for good dialogue, but also relevance, clarity, informativeness, benevolence and transparency (Miehling et al., 2024).

The importance of politeness is to connect language choice to interpersonal relations, particularly because politeness is important in this context. Brown and Levinson (1987) frame interaction in terms of a need for positive face (need for affirmation and affiliation) and negative face (need for autonomy and freedom from impingement). When a face-threatening act involves asking, refusing, criticism, complaint, or disagreement, it may be achieved through bald-on-record, positive-politeness, negative-politeness or off-record strategies. They also hypothesize that speakers consider social distance, relative power, and rank of imposition when deciding how much and what kind of redressive action to take. While heavily constrained by what is known, this framework can provide a clear analytic starting point for comparison of patterns of directness, mitigation, deference, and patterns of supportive language.

Speech-act studies show that pragmatic appropriateness does not necessarily follow from directness. The Cross-Cultural Speech Act Realization Project (CCSARP) identified three types of request strategies: direct, conventionally indirect, and non-conventionally indirect and found that there was quite a bit of variation in internal and external modification (Blum-Kulka & Olshtain, 1984; Blum-Kulka et al., 1989). Semantic formulas, explanations, gratitude, apology, and mitigation, too, are all part of the system of refusal, apology, and other face-sensitive acts (Beebe et al., 1990; Olshtain & Cohen, 1983). Natural interaction may be used between close interlocutors or in situations of one's role-based entitlement, while in a distant or upward status relationship it might be unnatural. So pragmatic competence refers to the calibration of one or more politeness markers in the context rather than maximization.

This is now becoming more important with conversational AI. Recent research indicates that LLMs can produce socially recognizable politeness but are not necessarily capable of reproducing the entire distribution of human choices. Zhao and Hawkins (2025) compared the production of humans and LLMs and identified some salient contextual politeness effects, while LLMs were overrepresented in using negative politeness strategies. Park et al. (2025) also found over empathy tendencies and a lack of clear relational cues in Korean DCT evaluation. In another discourse domain, Song et al. (2025) demonstrated that AI-produced counterspeech could still be distinguished from human-produced counterspeech and was found to be different with respect to politeness and specificity. These findings alone are sufficient to provoke analysis of how much and how the degree of politeness changes across social situations in an LLM, and if it does, how.

The present study aims to answer this question by conducting a controlled comparison of the Arabic DCT responses of human participants and ChatGPT. The original study was created using the same communicative situations and the same speech acts and contextual information were shared for both response sources. Because there are not directly comparable distributions of power, distance, and

imposition in naturally occurring human and AI conversations. Simultaneously, DCT evidence should be viewed as elicited pragmatic production and not as a faithful reflection of spontaneous DCTs, because written DCTs can vary in sequential organization, modification and interactional detail from naturally occurring DCTs (Economidou-Kogetsidis, 2013; Golato, 2003).

The study thus compares and contrasts the distribution of request strategies and supportive politeness devices across scenarios with the human participants, as well as the human modal realization to the realization produced by ChatGPT, and asks three questions: (1) how humans distribute their request strategies and supportive politeness devices across scenarios; (2) where the realization of the human participant overlaps with or deviates from the realization produced by ChatGPT; and (3) whether the AI output exhibits signs of contextual calibration or a more consistent tendency toward indirectness and mitigation. It is both empirical and methodological; it brings Arabic evidence for the pragmatics of human–AI interaction, and it explicitly distinguishes between descriptive comparisons between AI and human interaction and inferential claims, which would necessitate multiple independent generations of the models.

**Analytical Framework and Related Research**

The analysis is based on Brown and Levinson's (1987) model as a categorizing and interpretation model, instead of a literal criterion of politeness. Bald-on-record strategies involve the action itself; positive-politeness strategies focus on solidarity and/or approval; negative-politeness strategies involve minimizing imposition through indirectness, hedging, apology, deference, or optionality; and off-record strategies leave something of the intended meaning to inference. The present study is specifically interested in a resource being commensurate with the weight an act has in its context, rather than just existing or not.

The analysis is also based on the CCSARP tradition (Blum-Kulka & Olshtain, 1984; Blum-Kulka et al., 1989) for requests. The head acts are grouped into request head and external/supportive modification is coded separately. This is significant because two requests may have the same general head act strategy, but be very different in interpersonal work. A conventionally indirect request can take the form of a brief question setting up a request, while another can include an apology, grounder, formal address, positive preface, burden minimisation and the explicit opt-out clause.

The human–AI comparison is consequently evaluated at three levels. Categorical correspondence: Does ChatGPT make the same general realization that the typical human response makes? Pragmalinguistic correspondence relates to the question of whether similar linguistic devices occur. Contextual correspondence questions how much change occurs with the level of directness or mitigation in the scenario. This multi-layered analysis is not to equate surface politeness with pragmatic competence in human beings. An AI system doesn't have social face or interpersonal intentions in the human sense; the empirical question is whether there is a resemblance in the way in which language is produced to the contextually appropriate face management that humans engage in.

This also reflects recent studies on the socially oriented behaviour of an LLM. Zhao and Hawkins (2025) demonstrate that high aggregate performance does not necessarily mean systematic strategic biases, particularly higher use of distancing forms of politeness. Park et al. (2025) have shown that it is possible to have surface prosociality while at the same time being less sensitive to nuanced social relations. Similarly, Song et al. (2025) demonstrate that politeness can lead to observable differences between the outputs of AI and humans. Such studies indicate that human-likeness needs to be assessed at a distributionally and contextually appropriate level, but not from a single instance of fluent polite speech.

## METHOD

### Design and Data

The study employed a comparative mixed-method pragmatic research design with a quantitative approach (frequency analysis) and qualitative approach (interpretation). The human sample consisted of 44 participants who were randomly selected to answer 15 scenarios for the discourse completion task, which resulted in 660 valid responses (no missing data). Common requests mentioned in the scenarios were requests, gratitude/return, apology, disagreement, criticism/opinion, refusal, complaint, and advice. Eight scenarios resulted in requests, with 352 human request responses. In the source manuscript, the language used for the presentation of the DCT scenarios and for the answers of both the human participants and ChatGPT is highlighted as Arabic, meaning that the comparison is not on English proficiency, but rather on pragmatic realization.

The 15 scenarios were used for ChatGPT under the AI condition with the same generation context. The model was asked to answer naturally and briefly, in the situation described, as the speaker. For each scenario one response was generated to create an exploratory AI corpus of 15 responses comprising eight requests and seven responses of other speech acts. The source manuscript indicates that the model is ChatGPT (GPT-5.6 Thinking) and that it was accessed during the year of 2026 (OpenAI, 2026). Unlike the human sample, the AI corpus is considered to be a pilot sample for the AI model behaviour, as only one AI output was produced per scenario.

The scenarios were varied on the social distance, relative power, and degree of imposition, which are all related to Brown and Levinson's account of the weight of face-threatening acts. Situations were examples of closing down requests to close friends, professors, strangers, colleagues, and subordinate employees, apology, disagreement, criticism, refusal, complaint and advice situations. The setting of the DCT format, however, guaranteed the two sources of the responses to have identical situational information, despite its lack of the sequential and multi-modal characteristics of naturally occurring interaction (Economidou-Kogetsidis, 2013; Golato, 2003).

### Coding and Measures

Request head acts were broken down into one of four categories based on the CCSARP tradition: direct acts, conventionally indirect acts, hints, or non-target/non-performance. Greeting, solidarity address, formality, expressing politeness, apology, expressing gratitude/appreciation, reason/ground, opt out clause, minimizing the burden, compensation/reciprocity, and positive preface, were all coded as separate devices and mitigators.

The non-request speech acts were classified based on the classical speech-act perspectives (Olshtain & Cohen, 1983; Beebe et al., 1990), according to their pragmatic realization in the mode. The number of words for each human response and AI response was calculated for the length of the response.

The main thing that differentiates the human and AI comparison is that it's descriptive. The modal human realization for each scenario was contrasted with the realization produced by ChatGPT and alignment was coded where the broad categories matched. In the human sample, Cochran's Q was applied to determine if any difference occurred between the eight request types involving conventional indirect requests (Cochran, 1950). The length of the responses in each scenario was compared to the length of the single ChatGPT response by exploratory comparison, and a Wilcoxon signed-rank test was used. This is interpreted explicitly as exploratory as the two quantities have different sampling structures.

A null hypothesis was not tested to infer any difference in frequencies of human strategy vs. AI strategy at the population level. In fact, if the 15 AI outputs were considered to be 44 independent observations from models it would be pseudo-replication. Matched-proportion procedures (such as McNemar's test (McNemar, 1947)) and multiplicity adjustments (such as Holm's (Holm, 1979)) can be helpful in suitably replicated designs, but cannot be used for inference in the present AI corpus. Confirmatory testing would need to be done independently by repeated model generations.

## RESULTS

### Corpus Overview and Request Strategies

There were 660 valid human responses for 44 participants in the 15 scenarios. There were a total of 352 responses, across the eight request scenarios. Overall, there were more convention indirect requests (189/352, 53.7%) than direct requests (115/352, 32.7%), hints (27/352, 7.7%), or responses without a clearly identifiable target request (21/352, 6.0%). Significant differences existed between the 8 scenarios of conventional indirectness for human requests, Cochran's $Q(7)=60.66$, $p<.001$. The borrowing of a stranger's telephone and the request for a professor to repeat his/her spoken expression (84.1% and 79.5%, respectively) were associated with the highest prevalences, while the request for a close friend to provide large amounts of assistance (52.3%) and the request to an employee to revise a report (45.5%) were associated with the modal rates. It appears that the requests made by ChatGPT were more homogenous: of the 8 AI requests, all were of the conventional indirect query-preparatory type. This strategy was similar to the modal human realization in five of the six request scenarios, but was different in the scenarios where directness was the most common human strategy. It is therefore clear from the pattern that model always chose the low-risk indirect form even though human data showed a higher degree of relationally licensed directness.

**Table 1. Distribution of Human Request Strategies across Eight Scenarios**

| Scenario | Direct n (%) | Conv. indirect n (%) | Hint n (%) | No target n (%) | N |
|---|---|---|---|---|---|
| Close friend: substantial help | 23 (52.3) | 16 (36.4) | 2 (4.5) | 3 (6.8) | 44 |
| Late assignment acceptance | 18 (40.9) | 16 (36.4) | 10 (22.7) | 0 (0.0) | 44 |
| Professor: repeat utterance | 7 (15.9) | 35 (79.5) | 1 (2.3) | 1 (2.3) | 44 |
| Employee: revise report | 20 (45.5) | 16 (36.4) | 4 (9.1) | 4 (9.1) | 44 |
| Stranger's telephone | 3 (6.8) | 37 (84.1) | 2 (4.5) | 2 (4.5) | 44 |
| Project-deadline extension | 18 (40.9) | 19 (43.2) | 5 (11.4) | 2 (4.5) | 44 |
| Colleague: exchange shifts | 9 (20.5) | 30 (68.2) | 2 (4.5) | 3 (6.8) | 44 |
| Recommendation letter | 17 (38.6) | 20 (45.5) | 1 (2.3) | 6 (13.6) | 44 |
| Total | 115 (32.7) | 189 (53.7) | 27 (7.7) | 21 (6.0) | 352 |

***Note.*** Scenario percentages use N=44; total percentages use 352 human request responses.

**Supportive Moves and Mitigation**

The average number of coded supportive and/or mitigating devices per human request was 2.49. Highest occurred were grounders/reasons (65.1%), explicit politeness markers (58.2%), formal address (31.8%) and opt-out/conditional clauses (28.7%). ChatGPT had an average of 3.75 devices per request scenario. It contained a grounder in all 8 requests and in 62.5% it used opt-out clauses, 50.0% it used formal address, 50.0% it used apology, 37.5% it used positive prefatory remarks. The terms of the address were present in the human data but not in the AI requests. This descriptive pattern is consequently one in which there is more highly elaborated and standardised mitigation in the AI output, and more direct and brief human responses.

**Table 2. Supportive and Mitigating Devices in Human and ChatGPT Requests**

| Device | Human n | Human % | AI n | AI % | Pattern |
|---|---|---|---|---|---|
| Greeting | 18 | 5.1 | 1 | 12.5 | AI higher |
| Formal address | 112 | 31.8 | 4 | 50.0 | AI higher |
| Solidarity address | 58 | 16.5 | 0 | 0.0 | Human only |
| Politeness marker | 205 | 58.2 | 2 | 25.0 | Human higher |
| Apology | 57 | 16.2 | 4 | 50.0 | AI higher |

| | | | | | |
|---|---|---|---|---|---|
| Gratitude/appreciation | 40 | 11.4 | 1 | 12.5 | Similar |
| Grounder/reason | 229 | 65.1 | 8 | 100.0 | All AI requests |
| Opt-out clause | 101 | 28.7 | 5 | 62.5 | AI higher |
| Burden minimizer | 28 | 8.0 | 1 | 12.5 | AI slightly higher |
| Compensation/reciprocity | 14 | 4.0 | 1 | 12.5 | AI higher |
| Positive preface | 16 | 4.5 | 3 | 37.5 | AI higher |

***Note.*** Human percentages use 352 individual request responses; AI percentages use eight request scenarios. They are descriptive and are not equivalent inferential estimates.

**Non-request Speech Acts and Overall Correspondence**

For the gratitude/return, 97.7% of the human participants gave an explicit expression of thanks, while the modal realization was the expression of thanks alone (54.5%), which, in the case of ChatGPT, was paired with a return of the object. For the apology scenario, 97.7% of humans responded with an apology, with the most common being an apology formula (56.8%), while ChatGPT included explanation and concern. Disagreement with a manager was more heterogeneous: unmitigated, explicit disagreement accounted for 31.8% of the human responses, as did an indirect alternative, whereas ChatGPT employed a mitigated disagreement plus an alternative. Both sources preferred a ‘moderated criticism' or an ‘alternative' for criticism of an unfitting jacket by a close friend, the human modal category (45.5%).

The invitation scenario was highly modal with 88.6% of the responses being indirect refusals - ChatGPT was also highly modal in this area. In the restaurant scenario, a mitigated complaint and/or request for repair was modal (61.4%), and was also employed by ChatGPT. There were some differences in advice: Modal in humans (direct advice vs imperative): 65.9%, ChatGPT also gave an imperative, adding a reason and collaborative offer. The AI output matched the modal human realization in all 15 scenarios, with a 60.0% match rate.

The human responses had a mean of 10.47 words (SD = 7.68), while the responses from ChatGPT had a mean of 18.53 words (SD = 9.84). In all 15 cases the AI response was higher than the mean response for the human scenarios. The exploratory Wilcoxon signed rank comparison at the scenario level gave $W=0$, $p<.001$. This statistic should not be viewed as an estimate of ChatGPT's overall performance, but rather as evidence of the selected generation set for each AI value and the 44 responses used for each human value.

Table 3. Scenario-Level Comparison of Human and ChatGPT Pragmatic Realizations

| Sc. | Act | Human modal realization, n (%) | ChatGPT realization | Align. | Words H/G |
|---|---|---|---|---|---|
| 1 | Gratitude/ return | Thanks only, 24 (54.5) | Return + thanks | No | 4.16 / 8 |

| | | | | | |
|---|---|---|---|---|---|
| 2 | Request | Direct, 23 (52.3) | Conv. indirect | No | 11.77 / 25 |
| 3 | Request | Direct, 18 (40.9) | Conv. indirect | No | 12.34 / 20 |
| 4 | Request | Conv. indirect, 35 (79.5) | Conv. indirect | Yes | 7.48 / 12 |
| 5 | Request | Direct, 20 (45.5) | Conv. indirect | No | 12.57 / 21 |
| 6 | Apology | Apology only, 25 (56.8) | Apology + explanation + concern | No | 3.32 / 10 |
| 7 | Request | Conv. indirect, 37 (84.1) | Conv. indirect | Yes | 10.95 / 16 |
| 8 | Disagreement | Tie: explicit / indirect alt., 14 (31.8) | Mitigated + alternative | No | 9.70 / 20 |
| 9 | Criticism | Mitigated/alternative, 20 (45.5) | Mitigated + alternative | Yes | 7.84 / 12 |
| 10 | Request | Conv. indirect, 19 (43.2) | Conv. indirect | Yes | 14.00 / 29 |
| 11 | Refusal | Indirect mitigated, 39 (88.6) | Indirect mitigated | Yes | 11.18 / 19 |
| 12 | Complaint | Mitigated/repair, 27 (61.4) | Mitigated/repair | Yes | 9.86 / 18 |
| 13 | Request | Conv. indirect, 30 (68.2) | Conv. indirect | Yes | 11.75 / 20 |
| 14 | Advice | Direct/imperative, 29 (65.9) | Direct + collaborative offer | Yes | 13.68 / 17 |
| 15 | Request | Conv. indirect, 20 (45.5) | Conv. indirect | Yes | 16.39 / 31 |

***Note.*** Human modal percentages are based on 44 participants per scenario. H/G reports the human scenario mean and ChatGPT word count. ChatGPT contributed one generated response per scenario; alignment is descriptive.

## DISCUSSION

The cruising point is not merely a contrast between polite AI and less-polite individuals. Both sources provided practical and sensible answers, which were quite relevant in general. The more significant difference has to do with variability. The directness, conventional indirectness, hints, concise apology, mitigated refusal, complaint, and criticism were the ways human participants alternated in expressing apology.Human participants alternated between the following: directness, conventional indirectness, hints, concise apology, mitigated refusal, complaint, and criticism. ChatGPT, by contrast, had a more limited profile, characterized by more traditional indirectness, longer responses, and a greater number of mitigating statements supporting the claim. This is significant because while knowledge of polite forms is part of pragmatic competence, so is learning the amount of redress that is appropriate for a given relationship and act.

This is supported by the data related to requests made by humans. Overall, conventional indirectness was prevalent, but varied significantly by scenario. High rates of conventional indirectness were found when asking a stranger and a professor, which are in line with low entitlement, social distance, or institutional asymmetry. In contrast, directness was modal in the cases in which a speaker asked for significant assistance from a close friend and when an employee was asked to rewrite a report. The patterns illustrate the need to be able to distinguish the directness from impoliteness. Brevity can signal solidarity in familiar relationships; entitlement of the role can allow a direct directive in hierarchical work relationships. The utilitarian significance of form is thus a social question.

The use of conventional indirect forms of requests is impressive in ChatGPT, which is particularly impressive given its 100% usage.The strength and limitation of ChatGPT is that it is 100% conventionally indirect (with regard to request forms). The strategy was suitable and people friendly in a number of situations, particularly when dealing with strangers or institutional authority. Its persistence in contexts that were the human mode (direct requests) suggests overgeneralization, however. The model seems to be more open to a form that reduces the chance of overt face threat in the face of a more economical social realization. The result is similar to the findings of Zhao and Hawkins (2025), who found that models are particularly likely to use negative politeness strategies, and it can be aligned with the findings of Park et al. (2025), who found that models are more likely to produce prosocial or empathetic responses when humans can be more relationally differentiated.

Supportive-move density confirms the above interpretation. ChatGPT provided more coded supportive or mitigating devices per request and provided a grounder in each request. It also incorporated formal address, apologies and positive prefatory remarks more regularly than did the human participants. Together, these ensure a safe and explicit and respectful output, but when taken together can lead to unnecessary distance or elaboration. Similarly, Howe et al. (2023) found that ChatGPT's advice tended to be much longer than professional human advice, but this may be done without sacrificing the quality of the advice given. The present data bring the relevance of response elaboration to pragmatic speech-act production.

The non-request scenarios demonstrate that the model is not generally unable to make human-like selections. It correlated with the human modal category in mitigated criticism, refusal, complaint, advice, and in some request situations. The 60% statistic of this level of correspondence is of a pragmatic element. The divergences, however, are informative: When the human mode was shorter, ChatGPT tended to include explanation, concern, positive framing, and alternatives. This aligns with the general trend observed that AI-produced discourse can be discerned from human discourse in terms of politeness, specificity, and linguistic realization (Song et al., 2025). Therefore, a local fluency alone is not sufficient to judge human-likeness. The results also indicate that pragmalinguistic availability and sociopragmatic calibration are distinguishable. It is evident that ChatGPT can access linguistic forms of polite requesting, apologizing, refusal, complaint, and giving advice. The more challenging one is to choose from them according to social distance, power, imposition and relational familiarity. A model may be pragmalinguistically rich and at the same time be sociopragmatically regularized. To a human-AI design perspective, it does not mean the same to maximize conventional politeness as it does to maximize communicative appropriateness. This is pertinent to the general framework of conversation offered in the works of Miehling et al. (2024): the three dimensions of quantity, relevance, and manner relate to benevolence, and, socially effective output requires balancing interpersonal caution with economy and task efficiency.

Because pragmatic norms are not universal English templates, the Arabic context fills a gap that would otherwise be unfilled. Whether an utterance is respectful, intimate, distant, or formulaic can be determined by aspects of address, choice of dialect, degree of explicitness, and culturally interpreted entitlement. Please note that the present study does not provide any uniquely Arabic norms, since it does not compare the Arabic language with another language or dialect. The model, however, demonstrates that a system that could generate acceptable Arabic politeness could still regularize towards a more limited strategic repertoire than human participants to the same scenarios. This suggests the need for more multilingual work with multiple manipulation of social variables and repeated sampling of model outputs.

**Methodological Limitations and Implications**

The major constraint is the numerical inequality of the corpora. In a single standardized generation context, 44 humans responded to all 46 scenarios, while ChatGPT only responded to 1 scenario per scenario. The AI observations are then therefore not appropriate for making claims in terms of the distribution, variation or stability of model behaviour. This Wilcoxon result is to be treated as exploratory and should not be generalized to sessions, prompts or model updates. Future confirmatory work should produce a new independent output for each observation, preferably in the same number as the observations of humans in each scenario, and should record the exact snapshot of the model used, instructions given to the system, prompts given to the user, the sampling parameters, and the context of conversation.

Secondly, the DCT provides contextual control without conversational naturalness. It is not turn by turn negotiation, prosody, hesitation, repair, gesture, or developing common ground. Pragmatic method evidence demonstrates that the strategies used and the sequential organization of DCTs and naturally occurring interaction may vary (Economidou-Kogetsidis, 2013; Golato, 2003). Triangulation using role plays, recorded natural interaction, and multiple-turn human-AI dialogue would thus enhance the ecological validity.

Third, the scenarios are presented with both type of speech and sociopragmatic factors changed at the same time, making it impossible to attribute the difference on the basis of one single factor such as power, distance, or imposition. These dimensions should be varied independently of each other in a factorial design, with content of propositions being kept constant. Fourth, the analysis provided does not include an independent re-analysis of the coding phase. Before a confirmatory publication based on an expanded corpus, a second trained coder should classify a large portion of the responses and intercoder reliability should be expressed, such as Cohen's kappa (Cohen, 1960), and documented adjudication of differences.

Lastly, the production analysis does not determine the outputs' perception by the users. Depending on the listener and context, a response that is highly mitigated can be considered respectful, overly formal, distant, warm, unnatural, or efficient. Experiments of perception should thus be combined with assessment of politeness, naturalness, warmth, appropriateness, social distance and expected use rating, in order to complement the linguistic coding.

## CONCLUSION

The present study contrasted the politeness realization of 660 Arabic utterances from 44 human respondents with 15 responses by ChatGPT to the same discourse completion tasks. There was also a significant amount of contextual variation in human speakers, with indirectness being used when distance and authority meant it was a natural choice, but also in cases where brevity and directness were acceptable in relationships that allowed it. In nine of 15 scenarios, ChatGPT produced polite language that was more acceptable than human counterparts, but all of its eight requests were conventionally indirect, its responses were longer than the human ones, and its requests included more supportive mitigation. The contrast is best described as ‘human contextual variability vs. AI pragmatic regularization’. The results thus indicate a warning against the use of politeness-marker frequency as a measure for the assessment of socially responsive AI. A more pragmatically robust system needs to fine-tune directness, elaboration and mitigation to the social relationship and communicative load of the act. The AI corpus should be replicated, the sociopragmatic variables should be factored and manipulated, the models should be independently coded, and human perception ratings should be obtained in order to provide confirmatory evidence, as the present conclusions are exploratory.

## REFERENCES

Beebe, L. M., Takahashi, T., & Uliss-Weltz, R. (1990). Pragmatic transfer in ESL refusals. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), Developing communicative competence in a second language (pp. 55–73). Newbury House.

Blum-Kulka, S., House, J., & Kasper, G. (Eds.). (1989). Cross-cultural pragmatics: Requests and apologies. Ablex.

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). Applied Linguistics, 5(3), 196–213. https://doi.org/10.1093/applin/5.3.196

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

Cochran, W. G. (1950). The comparison of percentages in matched samples. Biometrika, 37(3–4), 256–266. https://doi.org/10.1093/biomet/37.3-4.256

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Economidou-Kogetsidis, M. (2013). Strategies, modification and perspective in native speakers' requests: A comparison of WDCT and naturally occurring requests. Journal of Pragmatics, 53, 21–38. https://doi.org/10.1016/j.pragma.2013.03.014

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. Applied Linguistics, 24(1), 90–121. https://doi.org/10.1093/applin/24.1.90

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. https://www.jstor.org/stable/4615733

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's advice is perceived as better than that of professional advice columnists. Frontiers in Psychology, 14, 1281255. https://doi.org/10.3389/fpsyg.2023.1281255

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157. https://doi.org/10.1007/BF02295996

Miehling, E., Nagireddy, M., Sattigeri, P., Daly, E. M., Piorkowski, D., & Richards, J. T. (2024). Language models in dialogue: Conversational maxims for human–AI interactions. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 14420–14437). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.843

Olshtain, E., & Cohen, A. D. (1983). Apology: A speech act set. In N. Wolfson & E. Judd (Eds.), Sociolinguistics and language acquisition (pp. 18–35). Newbury House.

OpenAI. (2026). ChatGPT (GPT-5.6 Thinking) [Large language model]. Retrieved July 29, 2026, from https://chatgpt.com/

Park, S., Kim, J., & Kim, H. (2025). Too polite to be human: Evaluating LLM empathy in Korean conversations via a DCT-based framework. In Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025) (pp. 76–89). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sicon-1.6

Song, X., Mamidisetty, S., Blanco, E., & Hong, L. (2025). Assessing the human likeness of AI-generated counterspeech. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 3547–3559). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.239/

Zhao, H., & Hawkins, R. D. (2025). Comparing human and LLM politeness strategies in free production. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 16188–16216). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.820