



Castledown

 OPEN ACCESS

Technology in Language Teaching & Learning

ISSN 2652-1687

<https://www.castledown.com/journals/tltl/>

Technology in Language Teaching & Learning, 6(3), 1–16 (2024)

<https://doi.org/10.29140/tltl.v6n3.1521>

Exploring the Potential of Artificial Intelligence for Instruction in Pragmatics



BRADFORD J. LEE^a 

DANIEL R. COOK^b 

^a*Fukui University of Technology, Japan*
bradford-lee@fukui-ut.ac.jp

^b*Fukui University of Technology, Japan*
cook@fukui-ut.ac.jp

Abstract

Large language models (LLMs) (e.g., ChatGPT) have garnered massive attention because of their potential to serve as interactive partners for a wide range of teaching and learning activities. In instructed second language acquisition, one potential use of LLMs is for teaching pragmatics. The current study sought to assess ChatGPT's suitability for this task by analyzing its output in response to 144 DCTs (discourse completion tasks), which test a language learner's ability to create apologies, requests, and refusals over different levels of social and power distance, and degrees of imposition. Quantitatively, ChatGPT produced apologies that were significantly shorter than the other two speech acts ($p < .001$, $R^2 = .49$), utilizing fewer pragmatic strategies ($p < .001$, $R^2 = .38$). Closer social distance and greater imposition also resulted in longer responses, with higher imposition also eliciting ChatGPT to use more pragmatic strategies. Qualitatively, human raters scored ChatGPT's apologies significantly lower in terms of their appropriateness, politeness, and language use, despite there being no statistical differences in the ratings for requests and refusals. The main pedagogical implications are that although ChatGPT offers an attractive option for students looking for more language exposure and interaction outside of the classroom, a degree of skepticism is still warranted because of its quantitative and qualitative inconsistencies.

Keywords: ChatGPT, large language models, pragmatics, pragmatic instruction, discourse completion tasks

Introduction

Students studying a second language (L2) typically focus their energies on mastering the vocabulary and grammar of the L2. However, significant attention should also be paid to developing *pragmatic*

Copyright: © 2024 Bradford J. Lee and Daniel R. Cook. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. **Data Availability Statement:** All relevant data are within this paper.

competence. After all, as the common phrase goes, many times it's not what you say, but how you say it. Unfortunately, many language students fail to master the L2's pragmatic cues, which can lead to mutual misunderstandings or the possibility of a complete communicative breakdown. Such a breakdown is disastrous, as not only do the students' confidence and motivation stand to be damaged, but their social image and standing may also be tarnished among the community if their speech comes across as being rude or uncouth, which, in turn, influences future interactions (Bardovi-Harlig et al., 1991).

Studies have shown that although some native speakers subconsciously alter their manner of processing when listening to non-native speakers because of the expectation of lower competence (Lev-Ari, 2015), this largely regards anticipating speech and the detection of lexical errors. Native speakers seem to be able to detect grammatical abnormalities in non-native speakers' speech and forgive them as errors, but not pragmatic missteps, which are typically misinterpreted as mere rudeness (Bardovi-Harlig et al., 1991). In fact, the more linguistically proficient the non-native speaker is, the higher the expectation of a congruent level of pragmatic competence seems to be. These findings indicate that a mastery of functional ability in the L2 is almost as important as the development of technical ability throughout the learning process, especially in the more advanced stages.

Unfortunately, numerous studies have demonstrated that in cases of English as a foreign language (EFL), where frequent exposure to authentic language is limited, pragmatic competence develops poorly without explicit instruction – that is to say, implicit instruction (i.e., natural acquisition as a result of exposure) is largely ineffective (Huang, 2022; Tajeddin & Bagherkazemi, 2021). Explicit instruction, on the other hand, is highly effective (e.g., Zhao, 2021). Despite these findings, however, pragmatic instruction is “often overlooked in the classroom and underrepresented in teaching materials and teacher education courses” (Siegel, 2016, p. 12).

The creation and widespread use of generative artificial intelligence (AI) has the potential to change the landscape of pragmatic instruction. As interactive AI chatbots become more adept at both processing and producing authentic language, the possibility of utilizing them for instruction in pragmatics becomes more viable, as teachers could potentially off-load pragmatics drills to be conducted via AI outside of the classroom. The current study was thus motivated by the need to assess the pragmatic competence of generative AI systems and thus their suitability to serve as models for pragmatic learners.

At the time of this writing, ChatGPT v3.5 is arguably the most market-dominant interactive AI system available to students and educators in the form of a web-interface and smartphone application. Being the first major large language model (LLM) to break onto the worldwide stage, ChatGPT benefits greatly from brand name recognition and boasts over 180 million users worldwide (Duarte, 2024). A newer version, ChatGPT v.4, is also currently on the market but requires a monthly subscription, which is somewhat prohibitive for casual users. As ChatGPT v.3.5 is available to all, it is the one most likely to be used by students and teachers, and is therefore of the greatest interest.

Review of the Literature

About ChatGPT

ChatGPT is a type of neural network with foundations in overlapping subfields of AI, as seen in Figure 1. GPT stands for Generative, Pre-trained Transformer, and such models are based on the transformer architecture developed by Vaswani et al. (2017) and have been implemented on a large scale by OpenAI. The large scale of ChatGPT and the techniques used for training it have had a profound result: a model that is very capable of a broad range of tasks despite not being specifically

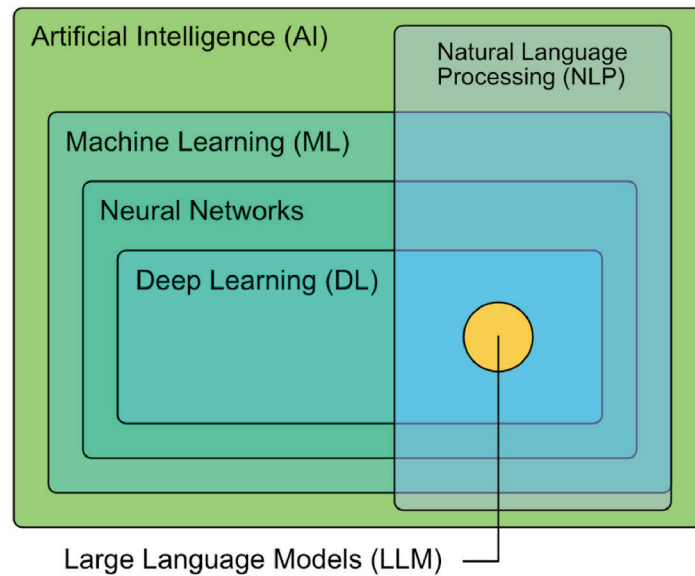


Figure 1 *Overlapping Subfields of Artificial Intelligence Comprising Large Language Models. Adapted from Goodfellow et al. (2016, p. 9).*

trained for them. The approach and techniques used for building ChatGPT have direct implications for its behavior. Thus, to assess ChatGPT, it is first essential to understand how it is made, how it works, and its general capabilities.

How ChatGPT is Made

OpenAI has not publicly released detailed technical information about the model (OpenAI, 2022). However, previous research on GPT models (Brown et al., 2020; Ouyang et al., 2022; Radford et al., 2018, 2019;), combined with technical overviews of the process (Minaee et al., 2024; Wu et al., 2023; Zhang, Zhang, et al., 2023) provide sufficient information. The following distills the pre-training, fine-tuning, and evaluation stages at a higher level from the abovementioned sources.

The first stage in building ChatGPT is pre-training. The goal of this stage is to produce a base model, which is done by exposing a model to a large dataset of unlabeled textual data. This process is called *unsupervised learning* and involves giving the model sequences of text, from which the model learns to make predictions. The resulting pre-trained model, used without modification, can perform well in several natural language processing (NLP) tasks (Brown et al., 2020). However, since the pre-trained model can have undesirable output, it is necessary to shape its behavior in a second stage of training.

The second stage is, therefore, fine-tuning. The goal of this stage is to shape the base model to have certain behaviors. OpenAI had two objectives for fine-tuning the base model (Ouyang et al., 2022). The first was aligning the base model with human intent by making the model “helpful, honest, and harmless” (Ouyang et al., 2022, p. 10). The second was to shape its behavior so that it behaved like an assistant. Fine-tuning was done in multiple steps with human feedback (Ouyang et al., 2022) in a type of fine-tuning called reinforcement learning with human feedback (RLHF).

Reinforcement learning exposes a model to rewards and punishments so that the model will modify its behavior to maximize rewards (Russell & Norvig, 2020, p. 789). In the case of ChatGPT, humans

provided examples of how the model should behave, as well as rating the responses that were used in the reward model (Ouyang et al., 2022; OpenAI, 2022). RLHF is performed through this repeated process: (1) a prompt is fed to a policy model to get a response, (2) the response is evaluated by a reward model, and (3) the evaluation is used as feedback to fine-tune both the policy model and the original pre-trained model.

The final stage for building ChatGPT is evaluation. A model is often trained to do specific tasks; it follows that evaluation using those tasks is necessary to measure the success of the training. ChatGPT was evaluated to see how well the model aligned with human intentions and its performance of NLP tasks (Ouyang et al., 2022).

It is notable that several factors involved in the design and creation of ChatGPT influence its behavior. The many components, methods, and techniques used in training can be seen as delicate and sensitive to changes. The architecture of the model, the type of training data, and the amount thereof will have an impact (Brown et al., 2020; Minaee et al., 2024; Radford et al., 2019), not to mention the finer design details. In addition, RLHF introduces reliance on human value judgements for several tasks (Ouyang et al., 2022, pp. 7, 19). Although steps are taken to minimize bias during pre-training and fine-tuning (Ouyang et al., 2022), the issues are not completely resolved and will certainly be reflected in its behavior.

How ChatGPT Works

The question of how ChatGPT works is difficult, as Bommasani et al. stated about LLMs generally: “[W]e currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties” (2022, p. 1). ChatGPT can be described mathematically as a series of calculations (Turner, 2024; Wolfram, 2023). However, a mathematical synopsis seems empty compared with the users’ experience: the responses it gives are human-like and dynamic.

What can be understood is what the neural network is designed to do:

“...what ChatGPT is always fundamentally trying to do is to produce a ‘reasonable continuation’ of whatever text it’s got so far, whereby [sic] ‘reasonable’ we mean ‘what one might expect someone to write after seeing what people have written on billions of webpages, etc.’” (Wolfram, 2023, p. 10)

There is no discovered theoretical reason for why ChatGPT works (Wolfram, 2023, p. 71), but Wolfram offered the perspective that “somehow in a neural net like ChatGPT’s it’s possible to capture the essence of what human brains manage to do in generating language” (Wolfram, 2023, p. 71).

At a high level, how ChatGPT works is simple: it receives a prompt, runs billions of calculations, and produces a sequence of slightly randomized predicted tokens as a response. The prompt is an important consideration when regarding behavior. When prompting ChatGPT, there are two parts sent to the model: a system prompt, which contains specific instructions that OpenAI gives to ChatGPT any time it is prompted, and a user prompt.

Related Work

Syntax, semantics, and pragmatics have been the chief concerns of NLP (Poole & Mackworth, 2023), and insights from broad surveys indicate that ChatGPT is adept at processing natural language to

assign meaning (Chang et al., 2023; Chang & Bergen, 2024; Laskar et al., 2023; Ziyu et al., 2023). Numerous studies have aimed to evaluate the social and pragmatic abilities of LLMs in a variety of ways; however, the terminology is not uniform despite overlap. We propose that socially focused studies and their related terms are better categorized under *social reasoning*. Social reasoning entails understanding the meaning of *what* is communicated or done and even *why* by inferring this from any number of social factors. Social reasoning is a precursor to and a necessity for *sociopragmatic knowledge*, which explores how language can be contextually applied in order to accomplish communicative goals. The interplay of social reasoning and sociopragmatic knowledge is *pragmatic competence*, a foundational aspect of communication. Our study is thus connected to a long history of NLP research, especially regarding evaluations of the social reasoning and pragmatic reasoning abilities of LLMs.

Social Reasoning Abilities

Research on social reasoning abilities is minimal, and studies in this area often group together and evaluate a range of LLMs, including ChatGPT. Evaluations include models that pre-date ChatGPT; however, given the similar architecture of LLMs, insights can be cautiously applied.

Sap et al. (2023) measured two domains related to participants in social interactions: social intelligence and Theory of Mind. Social intelligence refers to the ability of models to understand intentions and reactions, and Theory of Mind pertains to inferring the mental states of others and an ability to predict likely actions (Sap et al. 2023). They suggested that LLMs struggle to reason about social factors, despite performing well on other evaluations that involve pragmatic reasoning, though they noted that subsequent newer models (such as ChatGPT) have continued to show improvement. Choi et al. (2023) came to similar conclusions when they focused on evaluating the models' ability to infer meaning from five social categories: humor and sarcasm, offensiveness, sentiment and emotion, trustworthiness, and other social factors. Their conclusion that inferring social knowledge can be difficult for models is informative; however, some of the automated tests are misaligned between the tests' design and the purpose of the evaluation. Especially concerning is the exclusion of contextual information that is readily available in the original set of test data. This highlights a general issue with automated NLP evaluations: using very large datasets is not necessarily appropriate for evaluating LLMs for certain tasks, and new datasets and methods must be solicited.

Kocoń et al. (2023) evaluated ChatGPT in semantic and pragmatic categories with tests similar to those of Choi et al. (2023). They suggested that ChatGPT performs better semantically than pragmatically, specifically highlighting difficulties in reasoning about emotion. However, their tests for emotional sensitivity were based on single comments, stripped of context, making it impossible for the LLM to draw accurate inferences.

Sociopragmatic understanding was measured across 64 languages with automated evaluations by Zhang, Doan, et al. (2023). The results showed that ChatGPT does not have equal performance across languages, nor consistency across all measured categories. They reported that ChatGPT struggles with hateful or offensive speech towards a target and with dangerous, antisocial text. This is an obvious finding, since these issues are directly opposed to how ChatGPT is fine-tuned to be "helpful, honest, and harmless" (Ouyang, 2022, p. 10). A report on whether ChatGPT refused to respond to such prompts would be welcome.

From the above studies, we can tentatively infer that ChatGPT has general social reasoning abilities, but does not perform as well as humans. Currently, studies on ChatGPT's social reasoning abilities are insufficient to draw definitive conclusions, especially given the issues with methods and testing

approaches. Furthermore, given the limitations of current studies, it is possible that ChatGPT may have better performance but is not being measured accurately. We consider that approaching research in this area is complex and that evaluations need to be carefully designed. In addition, a deeper understanding of the limitations of social reasoning will require human evaluations. A severe limitation of ChatGPT is a lack of necessary perception: its text-only mode of communication entails constraints on the social factors it can utilize to infer meaning. Consequently, any social reasoning in communication is compromised. Thus, we note that ChatGPT has indeterminate weaknesses in this area. Further research is needed, not only for more concrete observations about social reasoning ability, but also to properly inform related topics like pragmatic abilities.

ChatGPT and other LLMs have become the state of the art in NLP, and as researchers uncover new capabilities and behaviors, new evaluations are in order. Thus far, the focus of evaluations has been on *reasoning* ability. These evaluations are important: without the ability to infer meaning from the linguistic and social elements included in the prompt, ChatGPT will not produce appropriate responses. ChatGPT has limitations, but its capabilities are sufficient to explore a variety of issues related to speech acts. Here, we explore evaluating the sociopragmatic knowledge of ChatGPT as a new area of study.

Research Questions

Stated formally, the current study sought to answer the following research questions:

1. Are there any quantitative differences in ChatGPT's output while performing the speech acts of requests, refusals, and apologies (i.e., response length, number of pragmatic strategies used, etc.)?
2. Are there any qualitative differences in ChatGPT's output to the above speech acts, as measured by human raters' judgements of its appropriateness, politeness, and language use?

Materials and Methods

Participants

In response to Research Question 2, which utilized human raters, a call was put out to the Pragmatics special interest group (SIG) of the Japanese Association of Language Teachers (JALT). Thirteen volunteers responded to these recruitment efforts. The self-reported cultural identities of the raters are reported below, in Table 1.

Table 1 *Self-Reported Cultural Identities of Human Ratets*

United States	6
United Kingdom	2
Australia	1
Germany	1
Canada	1
Jamaica	1
Japan	1

Materials and Methods

Research Question 1

To answer Research Question 1, which sought to assess ChatGPT's output when performing requests, refusals, and apologies, we used the three versions (Forms A, B, and C) of discourse completion task (DCT) Version #3, contained in Appendix E of Hudson et al.'s (1995) seminal work on cross-cultural pragmatics. DCTs work by setting up a scenario and asking the reader how they would respond linguistically. Within the description of each scenario, pragmatic information is explicitly prescribed: social distance (i.e., how familiar the interlocutors are with each other), power distance (i.e., if a hierarchical relationship exists between the interlocutors), and degree of imposition (i.e., the degree to which the scenario inconveniences each participant). This information is salient for completion of the exercise, as pragmatic choices depend on such contextual variables. Performance on DCTs is assessed on the basis of the number and type of pragmatic strategies employed in the responses (Hudson et al., 1995). Pragmatic strategies vary by speech type, but include up- and downgrades of language (e.g., *Hey Tom* < *Good morning, Mr. White*), acknowledgements (e.g., *I'm sorry I knocked over your glass...*), offers of correction (e.g., *I'll make it up to you*), among many others. While Hudson's development team (1995) balanced each DCT set to have an equal number of + and – scenarios (e.g., an equal number of scenarios where the speaker was more or less powerful than the listener), the current study used a more detailed approach and assigned relative weights to each scenario on a Likert scale of 1 to 9. As can be seen in Table 2, the variables were not perfectly balanced (i.e., a mean of 4.0), but skewed slightly towards having greater social and power distance, and degrees of imposition. Each version of the DCTs comprised 24 prompts (eight prompts for each of the three speech acts), resulting in 72 prompts in total (i.e., 24 for each speech act).

In order to reduce the “priming” effect, which modifies ChatGPT's behavior according to previous input in a chat session, a new session was started for each DCT. Each DCT was queried into ChatGPT via the following prompt:

Directions: Read the following situation then write what you would say in the situation in a normal conversation. Write “(nothing)” if you don't think the situation requires a response.

Situation: {situation text}

You say:

Each DCT was thus run through ChatGPT twice, resulting in 144 responses. ChatGPT did not refuse to answer any scenario. In 8 out of the 144 responses (5.56%), ChatGPT produced the exact same response twice in a row. In those cases, the duplicate response was discarded and the DCT was run a third time for the goal of having two unique responses per scenario (even a difference of a single word constituted a unique response).

Table 2 *Independent Variables Within Each Discourse Completion Task*

Variable	Mean	SD
Social distance	4.63	2.60
Power distance	5.92	3.17
Degree of imposition	4.58	2.95

Note. All variables: $n = 144$; possible range: 1–9.

Social distance: 1 = very close relationship; 9 = complete strangers.

Power distance: 1 = recipient is more powerful; 9 = speaker is more powerful.

Research Question 2

Research Question 2 investigated any potential qualitative differences in ChatGPT's output as measured by human raters' judgements of its appropriateness, politeness, and language use. Because of the excessive amount of data created in Research Question 1, only a single DCT set was selected for humans to rate, namely Set #1, which consisted of 24 DCTs, run two times each. The 24 prompts and 48 responses were compiled and entered into a Google Form. Instructions were provided, asking the raters to assess the responses to each scenario on a scale of 1 to 9, according to their appropriateness, politeness, and language use. Definitions of each dimension were also provided to ensure consistency across the raters (see Appendix A for full instructions). The ratings took approximately 30–45 minutes to perform.

Results

All variables and ratings were entered into a dataset using IBM's SPSS v.23 (Statistical Package for the Social Sciences), from which all statistical calculations were performed.

Research Question 1

First, descriptive statistics were extracted from the data, as reported below in Table 3.

In order to determine if the means reported in Table 2 were at the level of statistical significance, a multiple linear regression analysis was performed on the data, which indicated significant relationships of the independent variables with the word count [$F(4, 139) = 33.01, p < .001, R^2 = .49$] and also with the amount of strategies used [$F(4, 139) = 21.42, p < .001, R^2 = .38$]. Individual coefficients are reported in Table 4.

Table 3 *Descriptive Statistics*

Speech act	Word count, <i>M</i> (<i>SD</i>)	Strategies used, <i>M</i> (<i>SD</i>)
Apologies	25.15 (15.67)	3.38 (1.21)
Requests	39.46 (22.29)	3.88 (1.95)
Refusals	35.06 (15.72)	3.85 (1.07)

Note. For all speech acts, $n = 48$. *M*, mean; *SD*, standard deviation.

Table 4 *Output of Multiple Linear Regression*

		Beta	SE	β	<i>p</i>
Word count	Speech act	9.98	1.49	0.43	< .001
	Social distance	−1.61	0.49	−0.22	.001
	Power distance	−0.21	0.38	−0.04	.58
	Degree of imposition	4.71	0.43	0.73	< .001
Pragmatic strategies	Speech act	0.57	0.13	0.32	< .001
	Social distance	−0.02	0.04	−0.04	.61
	Power distance	−0.01	0.03	−0.03	.70
	Degree of imposition	0.32	0.04	0.64	<.001

To interpret the results reported in Tables 2 and 4, statistically significant relationships were discovered among the independent variables of speech acts, social distance, and degree of imposition on the word count of the responses that ChatGPT produced, while speech acts and degree of imposition were the only statistically significant relationships with the number of pragmatic strategies it employed. ChatGPT produced the longest responses when making requests (mean = 35.06), whereas the language it used for apologies was significantly shorter (mean = 25.15). ChatGPT also tended to produce more language when the social distance was closer (i.e., lower ratings of social distance) than when the interlocutors were judged to be further separated and when the degree of imposition was higher (i.e., higher ratings of imposition). With regard to the number of pragmatic strategies employed, while requests and refusals were both nearly equivalent at 3.88 and 3.85 on average, respectively, the number of strategies was significantly lower for apologies and scenarios with a lower degree of imposition.

Research Question 2

Next, the human raters' judgements of the appropriateness, politeness, and language use of ChatGPT's output ($n = 48$) were tabulated, with the descriptive statistics reported below in Table 5. Inter-rater reliability was also calculated for each set of variables, reported as Cronbach's alpha statistics.

As reported in Table 5, the Cronbach's alpha statistic across the entire dataset was $\alpha = .73$, which is generally considered to be an acceptable level of correlation (Tavakol & Dennick, 2011). However, the lowest inter-rater reliability was found for refusals, with the level of agreement for appropriateness, politeness, and language use being $\alpha = .10$, $.50$, and $.57$, respectively. The interpretation and implications of this will be presented in the Discussion section.

ANOVA (analysis of variance) was run on the data, as presented in Table 6, revealing multiple significant differences among the variables.

As reported in Tables 5 and 6, the appropriateness, politeness, and language use ratings of ChatGPT's apologies were all significantly lower than the ratings of the other two speech acts (i.e., requests and refusals). Conversely, the ratings of ChatGPT's requests and refusals were statistically equivalent across all speech acts.

Table 5 *Descriptive Statistics and Inter-Rater Reliability of Human Raters' Judgements*

Category	Speech act	M	SD	Cronbach's α
Appropriateness	Apologies	6.58	2.17	.71
	Requests	7.29	1.88	.69
	Refusals	7.56	1.78	.10
Politeness	Apologies	6.59	2.04	.81
	Requests	7.16	1.83	.62
	Refusals	7.59	1.59	.50
Language use	Apologies	6.87	2.00	.59
	Requests	7.39	1.65	.58
	Refusals	7.70	1.53	.57

Note. Overall Cronbach's alpha across the entire dataset: $\alpha = .73$. Possible Range: 1–9; M, mean; SD, standard deviation.

Table 6 *Output of Analysis of Variance*

DV	Speech act	Speech act	Mean difference	Std. error	p
Appropriateness	Apologies	Requests	−0.72	0.19	.001
		Refusals	−0.98	0.19	<.001
	Requests	Apologies	0.71	0.19	.001
		Refusals	−0.27	0.19	.34
	Refusals	Apologies	0.98	0.19	<.001
		Requests	0.27	0.19	.34
Politeness	Apologies	Requests	−0.58	0.18	.004
		Refusals	−1.01	0.18	<.001
	Requests	Apologies	0.58	0.18	.004
		Refusals	−0.43	0.18	.05
	Refusals	Apologies	1.01	0.18	<.001
		Requests	0.43	0.18	.05
Language use	Apologies	Requests	−0.53	.17	.006
		Refusals	−0.83	.17	<.001
	Requests	Apologies	0.53	.17	.006
		Refusals	−0.30	.17	.18
	Refusals	Apologies	0.83	.17	<.001
		Requests	0.30	.17	.18

Discussion

Research Question 1

The current study elucidated several tendencies of ChatGPT's performance when faced with situations requiring sociopragmatic knowledge, social reasoning, and ultimately pragmatic competence. The first result regarding Research Question 1 was that ChatGPT's output varied significantly depending on which speech act it was performing, indicating that ChatGPT can distinguish and act differently among (at least) apologies, requests, and refusals. Specifically, ChatGPT tended to produce shorter apologies than requests or refusals, and with fewer pragmatic strategies, as in these examples:

“Sorry for the wait, how can I assist you today?”

“I’m sorry, I’m in a bit of a rush to meet someone for lunch. Maybe another time, thank you!”

In the first example, the apology had a word count of 10 and a pragmatic strategy count of two (apology, offer of repair). The second example had a word count of 19 and a strategy count of four (apology, explanation, offer of future repair, expression of gratitude). Compare this with an example of a request and a refusal:

“Hey [lease-holder’s name], how’s it going? I wanted to ask if you could do me a favor this week. I’m heading out of town, and I usually take care of the lawn, but since I won’t

be here, would you mind mowing it this time? It usually takes me a couple of hours, so I'd really appreciate it if you could help out. Let me know if that works for you."

"I understand you're going out of town and need someone to cover your chores, but I'm swamped with work this week and won't be able to take on any extra tasks. Maybe we can work out a solution together; like finding someone else to cover for you or redistributing the chores among the rest of us?"

The request had 68 words and 10 strategies (lexical downgrade, greeting, preamble, explanation, self-promotion, explanation, request, explanation/self-promotion, expression of future appreciation, confirming agreement). Similarly, the refusal had a word count of 56, with six pragmatic strategies (acknowledgement of request, explanation, refusal, offer of alternative $\times 3$).

A possible explanation for this observed behavior may lie in ChatGPT's fine-tuning. In other words, it is possible that ChatGPT was fine-tuned to return quick and simple apologies in the context of its main purpose as a chatbot. Since ChatGPT is a user-facing product, OpenAI may have concerns regarding the users' experience, and a user can become dissatisfied with the product if they are given an unnecessarily verbose apology. Another consideration is financial, as OpenAI bears a cost per token, creating an incentive for shorter interactions. In addition, ChatGPT generally receives instructions and does not make requests from users, and refusals are often canned responses that are likely to be included in the system's prompt. OpenAI has hidden the mechanics of these issues from the public, but we may infer that any adjustments to ChatGPT with continual fine-tuning may have unintended consequences for practical applications of the model.

Research Question 2

The significantly shorter apologies found in Research Question 1 most likely contributed directly to the results elucidated in Research Question 2, namely that humans rated the appropriateness, politeness, and language use of ChatGPT's apologies significantly lower than the other two speech acts. However, this interpretation may not be entirely correct, as there is also the possibility that human raters downgraded the apologies for other reasons that are not discernable from the current data. For example, despite being instructed to return a response of "nothing" if the situation did not warrant a response, ChatGPT never went this route; it always produced language for all scenarios. It may be possible that humans reacted negatively to ChatGPT's apologies where the rater thought one was unwarranted. Future iterations of our research will therefore include exit interviews with the human raters in order to gain more insight as to the reasons for their quantitative up- or downgrades. Regardless, the current results suggest that ChatGPT's apology behavior is not well aligned with human expectations and shows a weakness of sociopragmatic knowledge in this regard.

One result to note was that regarding social distance and degree of imposition, ChatGPT produced much longer responses in cases of close social distance and a higher degree of imposition (and more pragmatic strategies were employed in cases of greater imposition). Both results fell in line with general expectations. As the degree of imposition grows, a larger volume of language is typically appropriate in order to convey to the listener the speaker's recognition of the higher stakes of the apology, request, or refusal. In the case of social distance, more language can typically be expected among interlocutors with a closer social distance, as the speaker is more invested in trying to ensure that the relationship with the listener is not negatively impacted, as opposed to interactions with unfamiliar interlocutors.

Finally, it is necessary to discuss the finding that the inter-rater reliability was extremely low regarding the appropriateness of ChatGPT's refusal strategies ($\alpha = .10$; Table 5), despite being relatively high ($.50 < \alpha < .81$) across all other variables. While the possibility exists that the variable was ambiguous

or the rating criteria were inadequately defined, the fact that the appropriateness ratings of apologies and requests were acceptable ($\alpha = .71$ and $.69$, respectively) somewhat mitigates this concern. Rather, the most likely explanation for this result is that it can be attributed to the subjectivity and complexity of refusals as a construct itself. Refusals are fundamentally different from apologies and requests in that they are made in response to a stimulus instigated by another party. Apologies and requests, on the other hand, are initiated by the speaker: they are active rather than reactive. Essentially, a person chooses whether they will apologize or not, or may change their mind and abandon making a request. However, when a request is made of us, the only choice is to reply (or ignore the speaker). Our results heavily imply that the reactive nature of refusals introduces a degree of personal bias and subjectivity into the length and degree of politeness, and even the need for a refusal at all for any particular circumstance. It should be noted, however, that despite the low Cronbach's α , the mean appropriateness score was 7.56 ($SD = 1.78$) (Table 5), suggesting that for refusals, there may have been more than a single underlying latent trait being measured.

Pedagogical Implications

The results of Research Question 2 (Table 5), suggesting that ChatGPT 3.5 makes generally appropriate apologies, requests, and refusals, implies that it can be a useful tool for language learners. For example, students could complete DCT assignments outside of class and then compare their answers with what ChatGPT produces as a model of an appropriate response. However, it appears that ChatGPT is not as good at apologizing as the other two speech acts, according to our data that its apologies are both shorter (Table 4) and were rated as less appropriate (Table 5). Nevertheless, we do not believe this statistical difference poses much of a practical difference, seeing as the appropriateness of its apologies was still rather high (a mean of 6.58), compared with means of 7.29 and 7.56 for requests and refusals, respectively. However, we emphasize that if users intend to use ChatGPT in this fashion, the utmost care must be taken in the way the prompts are composed. As ChatGPT is highly sensitive to the prompts' wording, the instructions, and prior interactions, we recommend that users adhere strictly to the guidelines we presented in the current study (Methods).

Limitations and Future Directions

While great care was taken to be deliberate and consistent at every stage, the constantly evolving state of ChatGPT and the obscurity behind its inner workings resulted in several limitations in our study. Firstly, the data collected from ChatGPT are not replicable, as the model is generative by design and includes randomization at a certain stage. In addition, our analysis included issues about the length of responses, which can be influenced by a number of internal factors that OpenAI can adjust without the user's knowledge, one of which is the continual fine-tuning of the model.

One possibility that cannot be ruled out is that ChatGPT was exposed to the DCT texts employed in the current study during its pre-training stage, as they were published in 1995 and, as seminal work, have been used extensively by linguistics researchers. The results could have been further contaminated if the DCTs had been coupled with responses, as in the case of a research manuscript or a published pragmatics lesson. However, since OpenAI is not open about the training datasets, this issue is not verifiable. Furthermore, the problem of data contamination in training datasets is more applicable to evaluations that require binary or scalar responses or a response from a provided list. In contrast, our experiment focuses on the text generated by ChatGPT and thus mitigating this concern somewhat.

Evaluations of models in NLP are often done with thousands of responses to generalize the data; however, this requires automation. Our method only collected two responses from ChatGPT for each DCT, and the results may not be generalized enough. Similarly, the range of social situations for DCTs and

the number of DCTs was limited. The DCTs have some similarity in their features which could help with generalization; however, future research will benefit from a greater quantity of both DCTs and gathered responses.

Related to this issue of quantity, some applications that require sociopragmatic knowledge may need a wide variety of responses from the model. A future direction could be to gather a variety of responses and assessing whether they are coherent with the situation and the range of variety (i.e., whether the full range of politeness levels is present).

Another limitation of the current study was the relatively small number of human raters (i.e., 13), of whom a large number (six) were from the United States. In order to garner more generalizable ratings, a larger rater pool with a wider distribution of the language communities that are represented will be gathered for future iterations of the study.

We have evaluated ChatGPT, yet there are other LLMs available publicly, and we anticipate more in the future. Evaluation of open-source models may provide further insight into the issues of incentives and risk avoidance for model training. Our methods relied on human evaluations, which presented issues. The potential for bias is present; moreover, human evaluations introduce financial and temporal requirements. A future direction could be to develop automated evaluations that incorporate assessments of the generated output, but they must be carefully designed to encompass the complexities involved.

Finally, the current study only analyzed the word count and number of pragmatic strategies that ChatGPT employed in response to the DCT stimuli. However, a more fine-grained analysis of the linguistic features of the responses would be welcome. For example, the current study noted changes in language such as *politeness markers* (e.g., Mr., sir, etc.) and *intensifiers* (e.g., thank you < thank you *very much*, etc.), but patterns and variations of syntax and grammar may reveal further patterns and/or differences between ChatGPT's behavior depending on the speech act.

Conclusion

The current study sought to investigate quantitative differences in ChatGPT's linguistic output for three different speech acts (apologies, requests, and refusals) by prompting it with 144 DCTs and analyzing the word count and pragmatic strategies in the responses (Research Question 1). The results indicated differential behavior based on the speech act: (1) apologies were significantly shorter and employed fewer pragmatic strategies, (2) the output was longer with closer social distance and a greater degree of imposition, and (3) more pragmatic strategies were used as the degree of imposition increased. This is evidence that ChatGPT can reason accurately, recognizing changes in the sociopragmatic conditions and adjusting its output. However, some of the tendencies in the output, such as the fact that its apologies were rated significantly lower in terms of their appropriateness, politeness, and language use, indicated that ChatGPT's fine-tuning is either insufficient or has resulted in unanticipated effects. In contrast to the typical automated methods employed in NLP that do not analyze output, our results provide two valuable contributions: novel evidence that can help practitioners to further improve models and a blueprint for further analyses going forward.

References

- Bardovi-Harlig, K., Hartford, B.A.S., Mahan-Taylor, R., Morgan, M.J., & Reynolds, D.W. (1991). Developing pragmatics awareness: closing the conversation. *ELT Journal*, 45(1), 4–15. <https://doi.org/10.1093/elt/45.1.4>

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the opportunities and risks of foundation models*. arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners*. arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). *A survey on evaluation of large language models*. arXiv. <https://doi.org/10.48550/arXiv.2307.03109>
- Chang, T. A., & Bergen, B. K. (2024). Language model behavior: a comprehensive survey. *Computational Linguistics*, 50(1), 1–58. https://doi.org/10.1162/coli_a_00492
- Choi, M., Pei, J., Kumar, S., Shu, C., & Jurgens, D. (2023). Do LLMs understand social knowledge? Evaluating the sociability of large language models with SockKET benchmark. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 11370–11403). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.699>
- Duarte, F. (2024, March 27). *Number of ChatGPT users (Apr 2024)*. Exploding Topics. <https://explodingtopics.com/blog/chatgpt-users>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Gubelmann, R., Kalouli, A., Niklaus, C., & Handschuh, S. (2023). When truth matters—addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). In A. Palmer & J. Camacho-collados (Eds.), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)* (pp. 24–39). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.starsem-1.4>
- Huang, N. (2022). Revisiting L2 pragmatic competence through implicit vs. explicit instructional framework. *Frontiers in Psychology*, 13(987729), 1–8. <http://doi.org/10.3389/fpsyg.2022.987729>
- Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Technical Report #7. Honolulu: University of Hawai'i at Manoa, National Foreign Language Resource Center.
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: jack of all trades, master of none. *Information Fusion*, 99, 101861. <https://doi.org/10.1016/j.inffus.2023.101861>
- Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). *A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets*. arXiv. <https://doi.org/10.48550/arXiv.2305.18486>
- Lev-Ari, S. (2015). Comprehending non-native speakers: theory and evidence for adjustment in manner or processing. *Frontiers in Psychology*, 5(21), 1546. <https://psycnet.apa.org/doi/10.3389/fpsyg.2014.01546>
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: a survey*. arXiv. <https://doi.org/10.48550/arXiv.2402.06196>
- OpenAI. (2022, November 30). *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>

- Poole, D. L., & Mackworth, A. K. (2023). *Artificial intelligence: foundations of computational agents* (3rd ed.). Cambridge University Press. <https://doi.org/10.1017/9781009258227>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: a modern approach* (4th edition). Pearson.
- Sap, M., LeBras, R., Fried, D., & Choi, Y. (2023). *Neural Theory-of-Mind? On the limits of social intelligence in large LMs*. arXiv. <https://doi.org/10.48550/arXiv.2210.13312>
- Siegel, J. (2016). Pragmatic activities for the speaking classroom. *English Teaching Forum*, 54(1), 12–19.
- Tajeddin, Z., & Bagherkazemi, M. (2021). Implicit and explicit pragmatic learning strategies: their factorial structure and relationship with speech act knowledge. *TESL-EJ*, 25(3), 1–28.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Turner, R. E. (2024). *An introduction to transformers*. arXiv. <https://doi.org/10.48550/arXiv.2304.10557>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- Wolfram, S. (2023). *What Is ChatGPT Doing ... And Why Does It Work?* Wolfram Media Inc.
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. <https://doi.org/10.1109/JAS.2023.123618>
- Zhang, C., Doan, K. D., Liao, Q., & Abdul-Mageed, M. (2023). *The skipped beat: a study of sociopragmatic understanding in LLMs for 64 languages*. arXiv. <https://doi.org/10.48550/arXiv.2310.14557>
- Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., Dam, S. K., Thwal, C. M., Tun, Y. L., Huy, L. L., kim, D., Bae, S.-H., Lee, L.-H., Yang, Y., Shen, H. T., Kweon, I. S., & Hong, C. S. (2023). *A complete survey on generative AI (AIGC): is ChatGPT from GPT-4 to GPT-5 all you need?* arXiv. <https://doi.org/10.48550/arXiv.2303.11717>
- Zhao, C. (2021). Explicit speech act instruction for developing EFL students' pragmatic competence. *International Journal of English Language Teaching*, 9(4), 26–35.
- Ziyu, Z., Qiguang, C., Longxuan, M., Mingda, L., Yi, H., Yushan, Q., Haopeng, B., Weinan, Z., & Liu, T. (2023). Through the lens of core competency: survey on evaluation of large language models. In J. Zhang (Ed.), *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)* (pp. 88–109). Chinese Information Processing Society of China.

Appendices

Appendix A – Instructions for Human Raters

Below are 24 individual scenarios, with two responses each (a total of 48 responses). Please read each scenario and the responses. Then assess the response, on a scale of 1–9, on its appropriateness, politeness, and language use (see end-cap descriptions below). Estimated time for completion: 20–30 mins.

Appropriateness

1. Very inappropriate request: The participant makes a request that is not relevant to the context or is inappropriate in some other way. For example, requesting time off without providing a reason or requesting time off for a reason that is not valid.

9. Extremely appropriate request: The participant makes a request that is exceptionally clear, relevant, and appropriate to the context, and demonstrates a very high level of language proficiency. For example, requesting time off with a clear and persuasive reason, providing alternative solutions, and using sophisticated language and tone.

Politeness

1. No politeness strategies: The participant does not use any politeness strategies in the request. For example, making a direct request without any polite language or expressions.

9. Very polite request: The participant uses very polite language or expressions, and demonstrates a high level of awareness of politeness conventions. For example, using formal greetings and closing expressions, using honorific language, and making indirect or hedged requests.

Language Use

1. Very poor language use: The participant uses incorrect or inappropriate language in the request, making it difficult to understand. For example, using incorrect grammar, vocabulary, or syntax.

9. Excellent language use: The participant uses excellent language in the request, with a sophisticated range of vocabulary and expressions, making it clear, engaging, and persuasive. For example, using complex sentence structures and expressions, and showing a high level of proficiency in formal language.