BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies
https://journals.bilpubgroup.com/index.php/fls

ARTICLE

AI's Struggle with Arabic: A Study on Pragmatic Failures in Contextual Communication

Shadi Majed Alshraah [1], ***Ashwaq Abdulrahman Aldaghri*** [2*], ***Iman Mohammad Oraif*** [2]

[1] *English Department, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia*

[2] *Department of English Language and Literature, College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 3204, Saudi Arabia*

ABSTRACT

AI systems, such as ChatGPT, often face challenges when using language in ways that fit social and cultural contexts, especially when making requests. While these models are strong in grammar and meaning, they frequently face challenges in capturing social and cultural aspects, leading to misunderstandings. To explore this issue, two frameworks were used: Taguchi's Pragmatic Appropriateness Model and Brown and Levinson's Politeness Theory. A mixed-methods approach compared AI and human responses to specific scenarios testing power, familiarity, and obligation. The findings reveal common problems, such as AI being overly formal in casual situations, misusing honorifics, mixing dialects, and misunderstanding context. These issues highlight the need for AI to better adapt to social and cultural differences, particularly in diverse environments. Integrating linguistic theories into AI training can enhance its ability to comprehend context and establish trust with users. This research stated that AI struggles to adapt to social norms, especially in situations where making requests requires accuracy. Most concerns involve being too formal, using honorifics incorrectly, mixing dialects in unnatural ways, and misunderstanding the context. These results highlight the need to include sociolinguistic principles in AI training to improve its understanding of culture and context. Furthermore, the results of the current study can help AI developers and policymakers in the MENA region.

*CORRESPONDING AUTHOR:
Ashwaq A. Aldaghri, Department of English Language and Literature, College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 3204, Saudi Arabia; Email: aaaldaghri@imamu.edu.sa

ARTICLE INFO
Received: 27 March 2025 | Revised: 17 April 2025 | Accepted: 10 September 2025 | Published Online: 29 October 2025
DOI: https://doi.org/10.30564/fls.v7i11.9252

CITATION
Alshraah, S.M., Aldaghri, A.A., Oraif, I.M., 2025. AI's Struggle with Arabic: A Study on Pragmatic Failures in Contextual Communication. Forum for Linguistic Studies. 7(11): 1537–1549. DOI: https://doi.org/10.30564/fls.v7i11.9252

COPYRIGHT
Copyright © 2025 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (https://creativecommons.org/licenses/by-nc/4.0/).

Highlights

- The paper identifies specific pragmatic failures encountered by AI systems when processing Arabic, highlighting issues such as misinterpretation of context, cultural nuances, and idiomatic expressions that affect communication effectiveness.
- Through a detailed analysis of various communication scenarios, the study demonstrates how AI's inability to grasp social cues and contextual subtleties leads to misunderstandings, thereby impacting user experience and trust in AI applications.
- The paper offers actionable recommendations for enhancing AI's performance in Arabic, including the integrating culturally relevant training datasets, improving natural language processing algorithms, and emphasizing the importance of human-in-the-loop systems to mitigate pragmatic errors.

Keywords: Artificial Intelligence (AI); Contextual Language Understanding; AI and Cultural Adaptation; Discourse Completion Test (DCT)

1. Introduction

Effective communication transcends grammatical precision, demanding pragmatic competence—the ability to adapt language to social contexts, relationships, and cultural norms. Humans intuitively calibrate requests based on variables such as power dynamics (e.g., addressing a superior versus a peer), familiarity (e.g., interacting with a stranger versus a friend), and obligation (e.g., asking for a favor versus fulfilling a duty). These adjustments ensure requests are perceived as appropriate, respectful, and contextually aligned. However, artificial intelligence (AI) systems, despite their proficiency in syntax and semantics, often falter in navigating these sociolinguistic subtleties. While models like ChatGPT generate grammatically flawless text, they frequently misjudge contextual norms, being overly polite in AI-generated requests, socially tone-deaf, or unnaturally phrased. Such pragmatic failures—instances where AI-generated language violates implicit social rules—undermine its utility in real-world applications, from customer service chatbots to educational tools, where nuanced communication is paramount [1,2].

The main challenge resides in reconciling linguistic theory with the practical demands of AI development. Human communication is inherently dynamic, shaped by unwritten rules of politeness, hierarchy, and cultural expectations. For example, a student requesting a deadline extension from a professor must employ deference and indirectness (e.g., "Would it be possible to grant me an additional day?"), whereas a peer might use brevity and informality (e.g., "Can you cover my shift?"). AI systems, however, often default to rigid or exaggerated strategies, such as overusing polite markers in casual contexts or misjudging power hierarchies [3,4]. These shortcomings stem from training data biases, a lack of contextual awareness, and insufficient integration of sociolinguistic frameworks into AI architectures.

This study investigates pragmatic failures in AI-generated requests through the lens of two foundational theories: Taguchi's Pragmatic Appropriateness Model, which emphasizes social variables (power, familiarity, obligation) as determinants of request appropriateness, and Brown and Levinson's Politeness Theory, which highlights context-dependent strategies like indirectness and hedging. Previous studies underscore AI's struggles in this domain: Nazcer et al. (2024) found ChatGPT misapplies indirectness in high-power scenarios, while Algouzi and Alzubi (2023) noted AI's tendency toward over-politeness in email replies, alienating users. These findings highlight a critical gap—AI lacks the sociolinguistic adaptability intrinsic to human communication. Given these challenges, this research aims to answer the following questions:

- How do AI-generated requests compare to human requests in terms of politeness, indirectness, and appropriateness across different social contexts?
- What are the most common pragmatic failures in AI-generated requests, and how do they vary based on social variables such as power, familiarity, and obligation?

To bridge this gap, a mixed-methods approach is employed, comparing AI-generated responses (ChatGPT) with human responses to Discourse Completion Test (DCT) scenarios that systematically vary power, familiarity, and obligation. By analyzing quantitative ratings (e.g., appropriateness, politeness) and qualitative feedback, prevalent failure types are identified, such as power misinterpretation or unnatural phrasing, and propose targeted solutions. It is important to understand these limitations to make AI better in situations where social awareness matters. For example, in education, an AI tutor should use language that is neither too casual nor too formal. In customer service, AI bots need to adjust to cultural differences to earn trust. To solve these issues, experts in language and AI should work together. By combining their knowledge, it is possible to create AI systems that are not only smart but also perfect at using language in ways that fit social and cultural situations.

2. Related Work

The study of AI's potential to use language in ways that fit social and cultural norms builds on years of linguistic theory and recent progress in AI research [5,6]. This section combines key ideas, studies on how AI communicates, challenges across cultures, and new methods to explain why this work focuses on areas where AI struggles with making appropriate requests.

2.1. Foundational Theories in Pragmatics and AI Communication

The theoretical foundations of pragmatic competence come from Taguchi's Pragmatic Appropriateness Model [7]. This model focuses on how social factors—like power, familiarity, and obligation—shape the way people use language in different situations. Taguchi's work, originally used to study second language learners, suggests that failing to follow these social rules leads to communication problems. This idea is now important for AI systems, which don't have natural social understanding.

Another key theory is Brown and Levinson's Politeness Theory [8]. It explains how people use strategies like being indirect, cautious, or respectful to avoid offending others, especially when making requests or giving criticism. The theory further demonstrates how politeness varies across cultures and relationships, which AI systems often struggle to handle.

Initial studies in computational pragmatics by Hovy [9] pointed out these challenges, noting that AI systems have trouble understanding unspoken social rules. This often leads to responses that feel stiff or out of place. Together, these theories provide a way to measure how well AI systems use language in social situations, especially when making requests where power and relationships matter.

2.2. Empirical Insights into AI's Pragmatic Shortcomings

Current studies show that AI still struggles to adapt its language to social situations. For example, Nazeer et al. [5] compared ChatGPT's responses to those of humans in request scenarios. They found that AI often uses overly direct language in situations where politeness is needed. For instance, instead of saying, "Could you please share the notes?" to a professor, ChatGPT might say, "Send me the notes." This happens because the AI's training data tends to favor straightforward, transactional language.

In the same vein, Algouzi and Alzubi [10] looked at AI-generated email replies in Gmail. They found that the AI often uses excessive politeness, like adding "Respected Sir/Madam" in casual conversations. This can make users feel that the language is fake or robotic.

Qiu et al. [11] also demonstrated that ChatGPT struggles to interpret implied meanings, like sarcasm or indirect refusals. For example, when someone says, "I'm swamped right now" to decline a request, the AI might respond with task suggestions instead of recognizing the refusal. These examples highlight how AI often relies on rigid or literal responses, missing the flexibility and tone of human communication.

2.3. Cross-Cultural and Contextual Challenges

AI's troubles with language are even more noticeable in cross-cultural situations, where politeness and power dynamics differ greatly. Cao et al. [12] studied how well ChatGPT follows social norms in 15 different cultures. They found major mismatches, like using overly direct language in places like Japan, where indirectness is valued, or being too formal in cultures like Sweden, where communication is more casual. For example, when simulating a worker's request to a manager in Japan, ChatGPT didn't use respectful language (like honorifics), making its responses seem rude.

Similarly, Paulikova [13] studied Slovak and Hungarian users interacting with chatbots. The AI often produced informal language in formal situations, like saying "Hi there!" during a professional healthcare conversation. These studies show that AI's training data, which tries to work for everyone, often fails to fit specific cultural contexts. To fix this, AI systems need datasets that include cultural details to improve their understanding of social and language differences.

2.4. Methodological Innovations in Evaluating AI Pragmatics

To identify and resolve AI's language issues, researchers have developed new methods. One common tool is the Discourse Completion Test (DCT), which has been used since Green's [14] work on human-robot interaction. DCTs create real-life situations, like asking a boss for a favor, to compare how humans and AI respond. This helps researchers study factors like power and familiarity.

Nam et al. [15] used Gricean maxims—such as being relevant and clear—to measure how AI responses differ from human ones. They found problems like AI being too wordy in situations where it was not necessary.

Guzman and Lewis [16] suggested adding user feedback to AI training. Their work revealed that people often notice small issues, like awkward phrasing or tone that automated systems miss. For example, users described AI requests as "stiff" or "too formal," even when the grammar was correct. These methods highlight the need for a mix of data-driven metrics and human feedback to fully understand and improve AI's ability to use language appropriately.

3. Methodology

3.1. Participants

The study involved 120 native-Arabic speakers, with 60 from Egypt and 60 from Saudi Arabia. All of participants are only men group and a range of ages (18–55 years) as **Figure 1** indicates. Egyptian participants spoke Cairene Arabic, using phrases like "ممكن تساعدني؟" ("Can you help me?"), while Saudi participants used regional dialects like Hijazi ("تقدر تساعدني؟" / "Can you help me?") or Najdi.

Participants were conducted from Prince Sattam University, as well as social media platforms like Facebook groups for Egyptian language learners and Saudi community forums. Regarding institutional ethical approval, both institutions required only the approval of the department head to conduct the study, as there was no risk of harm to the participants during the study. Based on the departments' approval, participants' consents were easily granted in both institutions. The researcher also explained the nature of the investigation to the participants, and their anonymity was ensured as no personal information of any of the participants was involved in the study.

The AI system used in the study was ChatGPT (GPT-4). It was set up to respond in Modern Standard Arabic (MSA) and regional dialects. The settings were adjusted for clear and natural conversation, with a severity of 0.7 and a limit of 150 tokens per response. For example, prompts specifically asked for responses in certain dialects, like "Write a request in Egyptian Arabic to borrow a book from a neighbor."

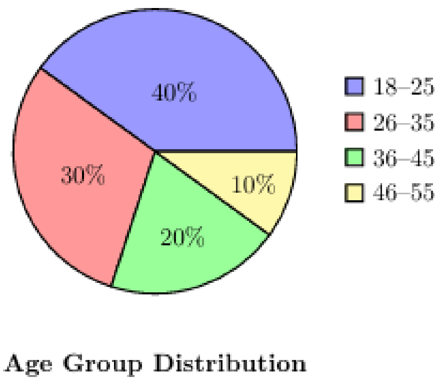

Figure 1. Demographic Distribution of Participants.

3.2. Data Collection

A Discourse Completion Test (DCT) with 12 scenarios was developed to reflect the social and cultural norms of Egypt and Saudi Arabia. The scenarios were designed to test three factors: power dynamics, familiarity, and obligation adpted by Alshraah, 2025

- **High-power scenarios** involved interactions with someone in authority, like asking a manager in Riyadh for a salary raise using Modern Standard Arabic (MSA).
- **Low-power scenarios** involved interactions between equals, like asking an Egyptian coworker in everyday Arabic to cover a work shift.
- **Familiarity** ranged from formal situations, like asking a Saudi bank teller in MSA to explain a transaction fee, to informal ones, like asking an Egyptian cousin in dialect to borrow a car.
- **High-obligation tasks** included important actions, like submitting a formal complaint to a landlord in Cairo using MSA.
- **Low-obligation scenarios** involved casual requests, like inviting a Saudi friend in Najdi Arabic to dinner.

Human participants filled out an online form, choosing to respond in MSA or their regional dialect. ChatGPT was provided the same prompts to generate responses. For example, the prompt “Write a request in Hijazi Arabic to reschedule a meeting with a colleague” resulted in AI responses like “عساك طيب، ممكن نغير الموعد لـ غداً؟” (“May you be well, could we reschedule for tomorrow?”).

3.3. Variables and Measures

The study looked at four main factors: power dynamics, familiarity, obligation, and language variety.

- **Power dynamics** were shown through role-based relationships. For example, in Saudi Arabia, people might use “حضرتك” (“Your presence”) when talking to a senior manager, while in Egypt, they might say “يا صاحبي” (“My friend”) to a peer.
- **Familiarity** was divided into interactions with strangers, like talking to Saudi customer service, or with friends and family, like asking an Egyptian sibling for help.
- **Obligation** distinguished important tasks, like work-related duties, from casual favors, like personal requests.
- **Language variety** was examined by comparing Modern Standard Arabic (MSA) with regional dialects. For example, Egyptian Arabic uses “عايز” (“I want”), while MSA uses “أرغب” (“I desire”).

The study also measured three outcomes: pragmatic appropriateness, politeness strategies, and naturalness.

- **Pragmatic appropriateness** checked if the language fit the culture. For example, in Saudi Arabia, people might use “شيخ” (“Sheikh”) for tribal leaders, while in Egypt, they might use “أستاذ” (“Mr.”).
- **Politeness strategies** analyzed at how indirect or polite the language was. For example, in Saudi Arabia, someone might say “عسى ما تكون زعلان” (“I hope you’re not upset”), while in Egypt, they might say “ممكن نعملها بكرة؟” (“Can we do it tomorrow?”).
- **Naturalness** measured how well the dialect fit the situation. For example, AI responses like “أرجو مساعدتك” (“I request your help”) in casual Egyptian settings were seen as too formal and unnatural.

3.4. Evaluation Framework

Quantitative analysis employed Likert-scale ratings (1–5), where 120 participants (60 Egyptian, 60 Saudi) analyzed AI and human responses on appropriateness, politeness, and naturalness. For example, participants rated statements such as “How respectful is this request to a Saudi manager?” with AI responses like \<أريد رفع راتبي> (“I want a raise”) scoring low (M = 2.1) in high-power Saudi scenarios due to bluntness. Three Arabic linguists further evaluated responses using a rubric based on Taguchi’s social variables (power, familiarity, obligation) and cultural norms, such as dialect accuracy and honorific usage. Qualitative analysis involved thematic coding of open-ended feedback, such as “The AI used MSA with my cousin, which felt weird,” categorizing failures into themes like over-formality (e.g., AI using \<الرجاء إعارة الكتاب> / “Please lend the book” in Egyptian informal contexts) or contextual missteps (e.g., Saudi AI using \<يا حاج> / “Pilgrim” in professional emails).

3.5. Data Analysis

Statistical analyses incorporated paired t-tests to compare AI and human mean scores, revealing significant gaps, such as AI scoring lower in Egyptian low-power scenarios (M = 2.8 vs. 4.3, $p < 0.001$). ANOVA tests assessed variance across cultural contexts, a p-value<0.05 was considered as an indication of statistical significance.

4. Evaluation and Results

4.1. Quantitative Analysis of AI vs. Human Responses

The study employed a mixed-methods approach to analyze AI-generated requests against human benchmarks. Likert-scale ratings (1–5) from 120 participants (60 Egyptian, 60 Saudi) revealed significant gaps in AI's pragmatic competence. For instance, in high-power scenarios, AI scored markedly lower than humans in appropriateness (M<sub>AI</sub> = 2.3 vs. M<sub>Human</sub> = 4.4, p < 0.001). In Saudi Arabia, AI's direct request \<أرسل لي التقرير> ("Send me the report") lacked deference compared to human phrasing \<هل يمكنك إرسال التقرير عند الإمكان؟> ("Could you send the report when possible?"). Similarly, AI overused polite markers in low-obligation contexts (M<sub>AI</sub> = 4.7 vs. M<sub>Human</sub> = 3.2, p < 0.01), for instance by producing overly formal language (\<أرجو مساعدتي> / "I kindly request assistance") for simple favors.

To evaluate the statistical significance of differences in AI performance across social variables,a one-way ANOVA was conducted. The analysis focused on three key variables: Power, Familiarity, and Obligation. The results are presented in **Table 1**.

Table 1. ANOVA Results for AI Performance Across Social Variables.

Social Variable	**F-value**	**p-value**
Power	8.9	<0.05
Familiarity	6.7	<0.01
Obligation	12.4	<0.001

The ANOVA results indicate that all three social variables significantly impact AI performance [17,18]. The highest F-value (12.4) and lowest p-value (<0.001) are associated with Obligation, suggesting that this variable has the strongest effect on AI's pragmatic competence. Power and Familiarity also had a significant impact, with F-values of 8.9 and 6.7. These results show how important it is for AI systems to understand and adapt to social hierarchies and relationships. The ANOVA results also revealed cultural differences:

- **In Egypt**, AI performed poorly in low-power, high-familiarity situations (F = 8.9, $p < 0.05$). For example, it struggled with dialect fluency, using formal MSA phrases like "هل يمكنك إعارة كتابك؟" ("Can you lend your book?") instead of the more natural dialect version humans used, like "تقدر تعيرني الكتاب؟" ("Can you lend me the book?").
- **In Saudi Arabia**, AI's politeness issues were most noticeable in high-obligation situations (F = 12.4, $p < 0.05$). For example, it failed to use honorifics when addressing superiors, saying "أرسل الملف الآن" ("Send the file now") instead of a more respectful phrase.

Figure 2 compares AI and human performance in three areas: Appropriateness, Politeness, and Naturalness. The results show clear differences in how they adapt to social norms:

- **AI struggled with Appropriateness (2.3)** and **Naturalness (2.5)**, meaning its requests often didn't fit social expectations and sounded stiff or overly formal.
- However, **AI scored high in Politeness (4.7)**, suggesting it overused polite language and indirectness, which can make its responses sound too formal or unnatural in casual situations.
- **Humans**, on the other hand, scored high in **Appropriateness (4.4)** and **Naturalness (4.6)**, while using politeness more carefully (3.2) based on the situation.

These findings show that AI struggles to adjust its tone and politeness dynamically, highlighting the need for better social and cultural adaptation in AI communication.

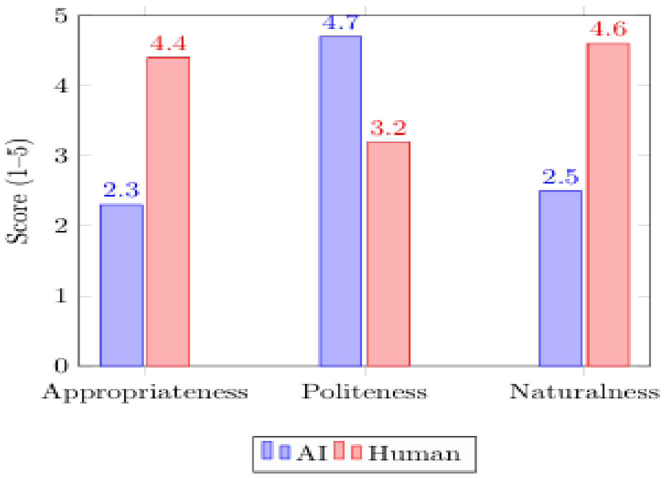

Figure 2. Comparison of AI and Human Performance in Key Metrics.

Figure 3 shows that AI's performance varies across different social situations, especially in high-power or unfamiliar interactions. For example, AI often misses the right level of respect and indirectness, making its responses too direct in formal settings (like talking to managers or professors) and too formal in casual ones (like talking to friends). While AI does better in low-pressure or familiar situations, it still falls short compared to humans, who consistently perform well in all contexts.

The findings indicate that AI struggles to adjust its politeness based on the situation, often misreading power dynamics and social relationships. To improve AI communication, future work should focus on helping AI better understand hierarchy, familiarity, and cultural norms. This will make AI interactions more appropriate and socially aware.

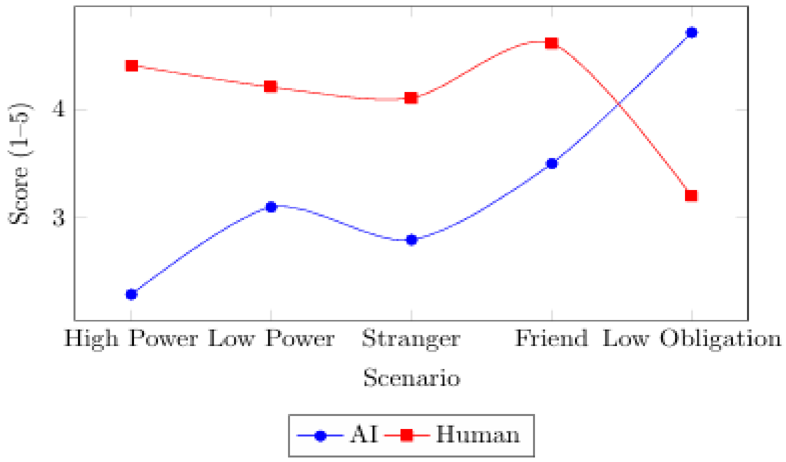

Figure 3. Performance Trends Across Social Scenarios.

4.2. Qualitative Insights: Recurring Failure Themes

4.2.1. Over-Formality in Casual Contexts

AI regularly defaulted to Modern Standard Arabic (MSA) in informal interactions, leading to perceptions of stiffness or robotic communication. For example:

Egyptian Context:

- AI Response: \<الرجاء إعادة المفتاح غدًا> (“Please return the key tomorrow”) in a casual request to a family member.

- Human Norm: \<رجع المفتاح بكرة يا عم> (“Bring the key back tomorrow, uncle”).
- Participant Feedback: “The AI sounded like a textbook—no one talks like this at home!”

Saudi Context:

- AI Response: \<سعادة المدير، أتمنى موافقتك> (“Honorable Manager, I hope for your approval”) when addressing a peer.
- Human Norm: \<يا محمد، ممكن توافق؟> (“Mohammed, can you agree?”).
- Participant Feedback: “Using ‘Honorable Manager’ with a coworker felt sarcastic.”

Implication: Over-reliance on formal MSA alienates users in casual settings, undermining trust in AI›s social adaptability.

4.2.2. Misplaced Honorifics and Titles

AI misapplied culturally specific honorifics, violating norms of deference and respect:

Egyptian Context:

- AI Mistake: Using \<يا أستاذ> (“Mr.”) for an elder family member.
- Expected Term: \<يا عم> (“Uncle”) or \<يا جدو> (“Grandpa”).
- Participant Feedback: “No one calls their grandpa ‘Mr.’—it felt cold.”

Saudi Context:

- AI Mistake: Addressing a senior professional as \<يا حاج> (“Pilgrim”), a term reserved for religious contexts.
- Expected Honorific: \<يا شيخ> (“Sheikh”) or \<سعادة الدكتور> (“Honorable Doctor”).
- Participant Feedback: “Calling my boss ‘Pilgrim’ was disrespectful and awkward.”

Implication: Misplaced honorifics signal a lack of cultural literacy, eroding user confidence in AI’s social competence.

4.2.3. Dialect Inconsistency and Mixing

AI struggled to maintain dialectal coherence, often blending MSA with regional dialects or using outdated phrases:

Egyptian Context:

- AI Response: \<أنا مشغول حالياً، ههزر معاك بعدين> (“I am busy currently, I’ll joke with you later”)—mixing MSA (\<حاليًا>) with Egyptian slang (\<ههزر>).
- Human Norm: \<مشغول دلوقتي، هكلمك بعدين> (“Busy now, I’ll talk to you later”).

Saudi Context:

- AI Response: \<تفضلوا بالجلوس> (“Please proceed to sit”), an archaic MSA phrase, in a casual Najdi context.
- Human Norm: \<اقعدوا> (“Sit”) or \<تَفضّل> (“Please”).
- Participant Feedback: “The AI sounded like a bad translator—half formal, half slang.”

Implication: Dialect mixing confuses users and reduces perceived fluency.

4.2.4. Contextual Misunderstanding

AI misinterpreted social cues, leading to tone-deaf requests:

Egyptian Context:

- AI Mistake: Framing a friendly favor as a demand: \<أحتاج إلى مساعدتك الآن> (“I need your help now”).
- Human Approach: Softening with humor: \<مش هتخلف عليا؟> (“Won’t you help me?”).

Saudi Context:

- AI Mistake: Using informal greetings (\<هلا> / “Hi”) in a formal business email.
- Human Norm: Starting with \<السلام عليكم> (“Peace be upon you”) and deferential language.
- Participant Feedback: “The AI didn’t understand when to be serious or playful.”

Implication: Contextual blindness makes AI appear socially inept.

Table 2 summarizes the failure types and explains their cultural impact.

Table 2. Summary Table: Qualitative Failure Types.

Failure Type	Example	Cultural Impact
Over-Formality	MSA in casual Egyptian requests (“Please return the key tomorrow”)	Perceived as robotic or insincere.
Misplaced Honorifics	Addressing a Saudi manager as “Pilgrim”	Seen as disrespectful or sarcastic.
Dialect Mixing	MSA + Egyptian slang (“I am busy currently, I’ll joke with you later”)	Confuses users; reduces fluency.
Contextual Missteps	Informal “Hi” in Saudi formal emails	Undermines professionalism.

4.3. Expert Evaluation Using Taguchi’s Rubric

4.3.1. Methodology and Rubric Design

This study compares AI-generated and human-generated requests through a structured scoring system based on Taguchi’s Pragmatic Appropriateness Model and cultural norms specific to Egypt and Saudi Arabia [19,20]. Three Arabic language experts—two specializing in Egyptian dialects and one in Saudi dialects—evaluated the responses. The scoring system looked at four main areas:

- **Power Sensitivity**: How well the language fits the power dynamic, like using respectful terms for superiors or casual language for peers.
- **Familiarity Adjustment**: Whether the language matches the level of closeness between people, like being formal with strangers or informal with friends.
- **Obligation Recognition**: How well the request balances politeness and directness, depending on how important the task is.
- **Dialect Accuracy**: Whether the AI uses Modern Standard Arabic (MSA) or regional dialects correctly, making the language sound natural in different situations.

4.3.2. Power Sensitivity: Hierarchical Failures

AI consistently misjudged power dynamics, particularly in high-authority scenarios.

- **Saudi Arabia**:
    - **AI Failure**: \<أرسل لي البيانات الآن> (“Send me the data now”) to a senior manager (score: 1.8/5).
    - **Human Benchmark**: \<هل يمكنك تزويدي بالبيانات عند الإمكان؟> (“Could you provide the data when possible?”) (Score: 4.9/5).
    - **Expert Comment**: “The AI ignored honorifics like \<سعادة المدير> (‘Honorable Manager’), essential in Saudi professional culture.”
- **Egypt**:
    - **AI Failure**: \<أحتاج إلى تقريرك اليوم> (“I need your report today”) to a professor (score: 2.0/5).
    - **Human Benchmark**: \<دكتور، ممكن التقرير يكون جاهز النهاردة؟> (“Doctor, could the report be ready today?”) (Score: 4.7/5).
    - **Expert Comment**: “The AI omitted \<دكتور> (‘Doctor’), a critical title for academic authority in Egypt.”

4.3.3. Familiarity Adjustment: Social Distance Missteps

AI struggled to adapt to familiarity levels, defaulting to formal language in casual contexts.

- **Egyptian Friend-to-Friend Requests**:
    - **AI Failure**: \<الرجاء إعارة كتابك> (“Please lend your book”) (score: 1.5/5).
    - **Human Benchmark**: \<تقدر تعيرني الكتاب؟> (“Can you lend me the book?”) (score: 4.8/5).
    - **Expert Comment**: Using MSA \<الرجاء> (‘please’) with friends is overly stiff—Egyptians prefer colloquial brevity.
- **Saudi Family Interactions**:
    - **AI Failure**: \<أتمنى مساعدتك في تنظيف المنزل> (“I hope for your help cleaning the house”) (score: 2.2/5).
    - **Human Benchmark**: \<يا أمي، ساعديني أنضف البيت> (“Mom, help me clean the house”) (score: 4.6/5).
    - **Expert Comment**: “The AI used distant, MSA phrasing instead of familial terms like \<يا أمي> (‘Mom’).”

4.3.4. Obligation Recognition: Over-Politeness and Directness

AI misbalanced politeness and directness, especially in low-obligation scenarios.

- **Egyptian Casual Favors**:
    - **AI Failure**: \<أرجو مساعدتي في إغلاق النافذة> ("I humbly request your assistance in closing the window") (score: 1.9/5).
    - **Human Benchmark**: \<ساعدني أغلِق الشباك> ("Help me close the window") (score: 4.5/5).
    - **Expert Comment**: The AI's exaggerated politeness (\<أرجو>) is inappropriate for trivial requests among peers.

- **Saudi Professional Follow-Ups**:
    - **AI Failure**: \<أحتاج إلى الرد فورًا> ("I need a reply immediately") (score: 2.1/5).
    - **Human Benchmark**: \<عسى ما تزعل، ممكن الرد بأقرب وقت؟> ("I hope you're not upset—could you reply soon?") (Score: 4.7/5).
    - **Expert Comment**: "The AI's directness (\<فورًا>) was perceived as aggressive—Saudi norms prefer softening phrases like \<عسى ما تزعل>."

4.3.5. Dialect Accuracy: MSA vs. Colloquial Fluency

AI's dialectal errors reduced perceived naturalness.

- **Egyptian Dialect Mixing**:
    - **AI Failure**: \<أنا مشغول حاليًا، ههزر معاك بعدين> ("I am busy currently, I'll joke with you later") (score: 2.0/5).
    - **Problem**: Mixed MSA (\<حاليًا>) with Egyptian slang (\<ههزر>).
    - **Human Benchmark**: \<مشغول دلوقتي، هكلمك بعدين> ("Busy now, I'll talk to you later") (score: 4.8/5).

- **Saudi Dialect Archaisms**:
    - **AI Failure**: \<تفضلوا بالرد> ("Please proceed to reply") (score: 2.3/5).
    - **Problem**: Used outdated MSA (\<تفضلوا>) in casual Najdi contexts.
    - **Human Benchmark**: \<ردوا علي إذا سمحتوا> ("Reply to me, please") (score: 4.6/5).

5. Discussion

This study underscores AI's continuing challenges with using language in socially appropriate ways, especially when making requests where social factors matter. While AI can create grammatically correct and polite responses, it often fails to match them to the right context, leading to misunderstandings [8]. This matches earlier research showing that AI struggles to understand and produce language that fits the situation [21].

A key challenge is AI being too formal in casual conversations. Instead of using regional dialects, AI often defaults to Modern Standard Arabic (MSA), making its responses sound stiff or robotic in everyday settings. For example, in Egyptian Arabic, AI used overly formal words that people usually save for professional situations, making casual conversations feel unnatural [22]. This shows AI's difficulty in adjusting its language style to different social settings [23].

A related issue is AI's misuse of honorifics and titles, which can come across as disrespectful or awkward. AI often fails to use the right titles based on cultural norms, like using religious terms in professional settings or leaving out expected titles in formal interactions. For instance, calling a senior Saudi manager "Pilgrim" instead of "Sheikh" made participants uncomfortable. This suggests AI lacks the cultural understanding needed to handle social hierarchies well [24]. AI also struggles with mixing dialects in unnatural ways. While it tries to use regional expressions, it often blends MSA with dialects poorly, creating responses that feel fake or confusing. Sometimes, AI even uses outdated phrases that people no longer use, making its language seem less authentic [22]. This shows the need for AI to better understand and use regional dialects correctly [23]. Additionally, AI often misreads power dynamics and obligation levels. In high-power situations, AI responses may lack the necessary respect, making them seem too direct or inappropriate. On the other hand, in low-power, casual situations, AI overuses politeness, making its responses sound overly formal. These mistakes suggest AI needs better training to understand factors like power, familiarity, and cultural expectations [8]. To address these issues, future AI models should include sociolinguistic frameworks that help them adapt their language to different contexts. This means using training data that covers a wider range of cultural and social situations and improving AI's ability to recognize and respond to contextual cues [24,25]. Adding interactive feedback systems could also help AI learn from real-time user interactions, making its responses more socially and culturally appropriate [7].

6. Conclusions

Our research uncovers a major gap between AI's ability to use correct grammar and its ability to use language in socially appropriate ways. The study found that AI struggles to adapt to social norms, especially in situations where making Arabic requests requires subtlety. In the case of the two main Arabic dialects, i.e., Egyptian and Saudi, key problems include being too formal, using honorifics incorrectly, mixing dialects in unnatural ways, and misunderstanding the context. These results highlight the need to include sociolinguistic principles in AI training to improve its understanding of culture and context.

Future work is needed to support AI in adjusting its language based on social hierarchies, power dynamics, and regional dialects. Combining AI development with expertise in linguistics will be key to creating AI systems that communicate more naturally and effectively across different cultures. By improving AI's ability to adapt to social situations, smarter systems that make human-computer interactions more meaningful can be built. Moreover, "future studies should add another group as native speakers of English which strengthen the results and use the data" [25].

Study Limitations and Recommendations for Future Studies:

1. This study is limited to only two informal Arabic dialects (Egyptian and Saudi); hence, future studies can examine other Arabic informal dialects.
2. The current study utilized ChatGPT as an AI tool, as it is the most popular system. Therefore, future studies can re-examine the use of similar Arabic texts employed in this study with other AI systems.

Author Contributions

S.M.A., A.A.A., and I.M.O. contributed equally to the conception, design, data collection, analysis, and writing of this study. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) [grant number IMSIU-DDRSP2504].

Institutional Review Board Statement

This study was conducted in accordance with ethical guidelines for research involving human participants. The research protocol was reviewed and approved by the relevant institutional ethics committee.

Informed Consent Statement

This study was conducted in accordance with ethical guidelines for research involving human participants. All participants provided informed consent before taking part in the study. Participants' anonymity and confidentiality were strictly maintained, and their responses were used solely for research purposes. No deceptive practices were employed, and all data collection adhered to ethical standards in linguistic and AI research.

Acknowledgment

The authors greatly appreciate the effort of every participant who took part in this study.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jakesch, M., Hancock, J.T., Naaman, M., 2023. Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences. 120(11), 2208839120. DOI: https://doi.org/10.1073/pnas.2208839120

[2] Warschauer, M., Tseng, W., Yim, S., et al., 2023. The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. Journal of Second Language Writing. 62, 1–20. DOI: https://doi.org/10.1016/j.jslw.2023.101071

[3] Zou, L., 2024. Creative computing in language understanding: A novel approach to pragmatic analysis [PhD Thesis]. University of Leicester: Leicester, UK.

[4] Hossain, K.I., 2024. Reviewing the role of culture in English language learning: Challenges and opportunities for educators. Social Sciences and Humanities Open. 9, 100781. DOI: https://doi.org/10.1016/j.ssaho.2023.100781

[5] Nazeer, I., Khan, N.M., Nawaz, A., et al., 2024. An experimental analysis of pragmatic competence in human-ChatGPT conversations. Pakistan Journal of Humanities and Social Sciences. 12(1), 424–435. DOI: http://dx.doi.org/10.52131/pjhss.2024.v12i1.2061

[6] Alos, J., 2015. Explicating the implicit: an empirical investigation into pragmatic competence in translator training. The Interpreter and Translator Trainer. 9(3), 287–305. DOI: https://doi.org/10.1080/1750399X.2015.1100398

[7] Taguchi, N., 2009. Pragmatic competence in Japanese as a second language: An introduction. Pragmatic Competence. 5, 1–18. DOI: https://doi.org/10.1515/9783110218558.1

[8] Brown, P., Levinson, S.C., 1987. Politeness: Some Universals in Language Usage. Cambridge University Press: Cambridge, UK.

[9] Hovy, E.H., 1990. Pragmatics and natural language generation. Artificial Intelligence. 43(2), 153–197. DOI: https://doi.org/10.1016/0004-3702(90)90084-D

[10] Algouzi, S., Alzubi, A.A.F., 2023. The study of AI-mediated communication and sociocultural language-related variables: Gmail reply suggestions. Applied Artificial Intelligence. 37(1), 2175114. DOI: https://doi.org/10.1080/08839514.2023.2175114

[11] Qiu, H., Zhao, T., Li, A., et al., 2023. A benchmark for understanding dialogue safety in mental health support. arXiv preprint. arXiv: 2307.16457v1. DOI: https://doi.org/10.48550/arXiv.2307.16457

[12] Cao, Y., Zhou, L., Lee, S., et al., 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. arXiv preprint. arXiv: 2303.17466. DOI: https://doi.org/10.48550/arXiv.2303.17466

[13] Paulikova, K., 2025. Navigating interaction with AI chatbots: Sociolinguistic and pragmatic aspects of chatbot use by Slovak and Hungarian speakers. Journal of Siberian Federal University. Humanities & Social Sciences. 18(1), 70–80.

[14] Green, A., 2009. Designing and evaluating human-robot communication: Informing design through analysis of user interaction [PhD Thesis]. KTH Royal Institute of Technology Stockholm, Sweden.

[15] Nam, Y., Chung, H., Hong, U., 2023. Language artificial intelligences' communicative performance quantified through the Gricean conversation theory. Cyberpsychology, Behavior, and Social Networking. 26(12), 919–923. DOI: https://doi.org/10.1089/cysber.2022.0356

[16] Guzman, A.L., Lewis, S.C., 2020. Artificial intelligence and communication: A human-machine communication research agenda. New Media and Society. 22(1), 70–86. DOI: https://doi.org/10.1177/1461444819858691

[17] Kim, H.-Y., 2014. Analysis of variance (ANOVA) comparing means of more than two groups. Restorative Dentistry and Endodontics. 39(1), 74–77. DOI: https://doi.org/10.5395/rde.2014.39.1.74

[18] Kim, T.K., 2017. Understanding one-way ANOVA using conceptual figures. Korean Journal of Anesthesiology. 70(1), 22–26. DOI: https://doi.org/10.4097/kjae.2017.70.1.22

[19] Moffett, J.W., Fennell, P., Harmeling, C.M., et al., 2024. The Taguchi approach to large-scale experimental designs: A powerful and efficient tool for advancing marketing theory and practice. Journal of the Academy of Marketing Science. 53, 949–954. DOI: http://dx.doi.org/10.1007/s11747-024-01059-0

[20] Taguchi, N., Crawford, W., Wetzel, D.Z., 2013. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. TESOL Quarterly. 47(2), 420–430.

[21] Devlin, J., Chang, M.-W., Lee, K., et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. DOI: https://doi.org/10.48550/arXiv.1810.04805

[22] Habash, N.Y., 2010. Introduction to Arabic natural language processing (Synthesis lectures on human language technologies). Machine Translation 24, 285–289. DOI: https://doi.org/10.1007/s10590-011-9087-8

[23] Zhang, Y., Sun, S., Galley, M., et al., 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv: 1911.00536. DOI: https://doi.org/10.48550/arXiv.1911.00536

[24] Blodgett, S.L., Barocas, S., Daume III, H., et al., 2005. Language (technology) is power: A critical survey of "bias" in NLP. arXiv preprint. arXiv: 2005.14050v2. DOI: https://doi.org/10.48550/arXiv.2005.14050

[25] Alshraah, S.M., Harun, H., Kariem, A.A., 2023. Pragmalinguistic competence of directness request level: A case of Saudi EFL learners. International Journal of Society, Culture & Language. 1–16. DOI: https://doi.org/10.22034/ijscl.2023.2009932.3139