Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05282-0 OPEN

Metaphor interpretation in Jordanian Arabic, Emirati Arabic and Classical Arabic: artificial intelligence vs. humans

Aseel Zibin[1,2]✉, Nabeeha Binhaidara[3], Hala Al-Shahwan[1] & Haneen Yousef[1]

This study examines how well humans, both Jordanians and Emiratis, and four AI tools—ChatGPT-4, ChatGPT-3.5, Google Gemini, and Ask PDF—can understand metaphors in Classical Arabic (CA) and its everyday forms in Jordanian Arabic (JA) and Emirati Arabic (EA). We tested fifty participants from Jordan and the UAE on their grasp of various colloquial and CA metaphorical expressions. Two distinct tests were employed, each comprising 40 items. Test 1 was administered to Jordanian participants and included 20 metaphorical expressions in Jordanian Arabic and 20 metaphorical expressions in Classical Arabic. Similarly, Test 2 was administered to Emirati participants and contained 20 expressions in Emirati Arabic and 20 expressions in Classical Arabic. The Mann–Whitney U test was employed to evaluate differences in accuracy and interpretation between AI tools and human participants from both regions in the contexts of colloquial and Classical Arabic. The results showed that participants from Jordan had a better understanding than the AI tools, likely due to their strong cultural background. In contrast, the Emirati participants performed similarly to the AI. The AI tools were more effective at interpreting CA metaphors compared to Emirati participants; AI tools are typically trained on diverse datasets and that usually leads to strong performance in interpreting formal or Classical Arabic expressions. These findings emphasize the need for improvements in AI models to boost their language processing abilities, as they often miss the cultural aspects required for accurately interpreting figurative language. This study adds to the ongoing discussion about AI and language interpretation, revealing both the potential and the obstacles AI faces when dealing with culturally rich and context-sensitive language.

[1] University of Jordan, Amman, Jordan. [2] Applied Science Private University, Amman, Jordan. [3] Mohammed Bin Zayed University for Humanities, ABU DHABI, United Arab Emirates. ✉email: a.zabin@ju.edu.jo

Introduction

Artificial Intelligence (AI) seeks to replicate human cognitive abilities; however, it encounters significant challenges in comprehending certain aspects of language, particularly metaphors. Unlike literal expressions, metaphors rely heavily on context, world knowledge, and cultural understanding—elements that static algorithms struggle to capture. Current AI systems often depend on keyword matching or pre-programmed schemas, neglecting the dynamic interaction of linguistic context, situational factors, and cultural knowledge essential for accurate interpretation (Boden 2016). Although some models, such as MIDAS (Martin 1990) and Fass' system (1997), employ techniques like conceptual metaphor mapping and analogy-based reasoning, they still fall short in explaining the richness and ambiguity of real-world metaphorical language. The unique linguistic diversity of Arabic, characterized by variations between Classical and colloquial forms, provides a compelling context for examining these limitations[1]. This study aims to address this gap by comparing the performance of human participants with four state-of-the-art AI tools—ChatGPT-4, ChatGPT 3.5, Google Gemini, and Ask PDF—in interpreting metaphors within Classical Arabic and its colloquial variants, specifically Jordanian and Emirati Arabic.

The selection of ChatGPT-4, ChatGPT 3.5, Google Gemini, and Ask PDF was based on several key criteria. These models exemplify a range of leading AI architectures and natural language processing (NLP) approaches, allowing for a comparison of their capabilities in metaphor interpretation. ChatGPT-4 and ChatGPT 3.5 are acknowledged for their robust performance in general language understanding, while Google Gemini was added to examine its innovative methods for interpreting textual data. Ask PDF was included for its capacity to analyze text from diverse sources, potentially offering a distinctive perspective on metaphor extraction. This varied selection facilitates a better evaluation of state-of-the-art models, emphasizing their strengths and weaknesses in handling metaphorical language. The choice of Jordanian and Emirati Arabic was made thoughtfully. Jordanian Arabic (JA), which is a Levantine dialect, and Emirati Arabic (EA), a Gulf dialect, together reflect the rich diversity found within the Arabic language. This selection allows us to look into how different dialect features influence AI's understanding of metaphors. Additionally, we considered practical factors such as data availability and access to native speakers. This balanced approach provided us with a clearer picture of how well AI performs across key Arabic dialect groups.

Enhancing AI's understanding of metaphorical language across different dialects could significantly improve cross-cultural communication and translation technologies. The accuracy of AI interpretations is assessed against human judgments, contributing to advancements in natural language processing. Specifically, this study investigates the following research questions:

(1) To what extent can AI tools (ChatGPT-4, ChatGPT 3.5, Google Gemini, and Ask PDF) accurately interpret metaphors in Classical, Jordanian, and Emirati Arabic?
(2) How do human interpretations compare to those produced by the AI tools across the three Arabic varieties?

Literature review

Theoretical framework

Conceptual metaphor theory as based on the idea of main meaning focus. The Conceptual Metaphor Theory (CMT) has faced several criticisms. One significant issue is schematicity: not all features of the source domain consistently map onto the target domain (see Zibin and Altakhaineh 2023). For example, in Lakoff and Johnson's (2003) metaphor, theories are buildings; a theory may have a "solid foundation," but describing it as having 'high windows' is nonsensical (Clausner and Croft 1997). Furthermore, CMT's emphasis on cultural differences conflicts with the embodied cognition perspective, which posits that metaphors are grounded in universal bodily experiences and image schemas (Kövecses 2008). This tension highlights the ongoing debate between the universality of certain metaphorical mappings and their culturally specific variations.

Given the weaknesses in standard CMT, several proposals have been suggested to enhance it. Kövecses (2003) argued that source domains are typically correlated with one basic meaning that differentiates them and provides uniqueness to each concept. This meaning comes from core knowledge about the specific linguistic item, which is then attributed to the target domain through mapping. The core knowledge is the conventional aspect of meaning (Langacker Ronald 1987).

In this view, meaning is considered the main knowledge, a prominent feature in communities regarding a certain concept that gets attributed to another. The meaning focus could be innately rooted in the word itself, e.g., the concept of building, or it may emerge when comparing one concept to another that is metaphorically correlated (Kövecses 2008). Meaning emerges against the background of another concept; for example, the main meaning focus of butchery is incompetence and sloppiness against the background of surgery, while the main meaning focus of surgery is precision and seriousness against butchery. Consequently, in "the surgeon is a butcher," the conventional knowledge of butchery, when compared to surgery, emerges as incompetence and sloppiness. However, in "my sister's criticism was a butchery," the main meaning focus of butchery, built on the concept of criticism, emerges as harsh and damaging, reflecting the common knowledge of brutal cutting.

Furthermore, in "Today's argument was a butchery of reason," the main meaning focus of butchery emerges as the end and death of logic in that specific argument, highlighting the aspect of killing associated with butchery. According to Kövecses (2011), this emergence results from the extension of meaning; the standard meaning of *butchery* as the act of slaughtering animals is extended to sloppiness and incompetence in the context of surgery. This theory is adopted in this study to interpret the answers of human participants against those of AI tools. The next section provides a general background of the study and reviews related studies.

Related studies. Although AI has gained attention since the emergence of ChatGPT, it existed long before. Earlier developments in machine learning and natural language processing laid the groundwork for sophisticated AI models like ChatGPT. Various systems have been used to process and interpret conceptual metaphors, including Hobbs et al. (1993) ATT-Meta system, Veale's (1998) Sapper model, Narayanan's metaphor-understanding system (1997, 1999), and Barnden's ATT-Meta approach (Barnden 2001; Barnden et al. 1994). Additionally, several studies have explored AI's ability to generate metaphorical expressions (Gargett and Barnden 2013; Gargett et al. 2015; Di Biagio 2022). However, there is a lack of studies demonstrating AI's effectiveness in interpreting metaphorical language. This review aims to establish a foundation for the current study.

Wachowiak and Gromann (2023) evaluated ChatGPT-3's ability to detect and understand conceptual metaphors by selecting 446 metaphors from Lakoff's Master Metaphor List. They included 50 non-metaphorical sentences from the VUA corpus to prevent the model from assuming every expression is metaphorical. The dataset was divided into training, validation, and test sets, ensuring the test set contained different expressions. They added 284 English sentences and 110 Spanish sentences from the LLC dataset to further challenge the model with longer and more complex sentences. Two training techniques were used: Few-Shot Prompting, which provides a few examples before new sentences, and Fine-Tuning, which uses a larger set of examples for deeper training. To evaluate the model's performance in the validation test, automated metrics were used, while the accuracy on the test set was assessed through manual review of ChatGPT-3's answers.

The results showed that ChatGPT-3 predicted the source domain with an average accuracy of 60.22% across three datasets, and an accuracy of 65.15% in English and 34.65% in Spanish. The decrease in the model's accuracy with Spanish metaphorical expressions may be due to differences in language complexity or GPT-3's training data coverage. Additionally, the model suffered from errors, particularly hallucinating domains unsupported by any clues in the sentence. It also struggled to identify metaphorical elements, treating them as literal, and predicted incorrect source domains based on words not metaphorically related to the target domain.

Tong et al. (2024) conducted a similar study to evaluate the capability of LLMs to understand metaphors, including ChatGPT-3.5 and LLaMA. The researchers designed the Metaphor Understanding Challenge Dataset (MUNCH), which contains over 10,000 paraphrases for metaphorical expressions and 1500 triples, each containing a metaphorical expression and two paraphrases—one apt and one inapt. These metaphorical expressions were sourced from the VUA corpus across four genres (academic, news, fiction, and conversation) and featured varying levels of novelty. The researchers set up two tasks for the LLMs to evaluate their comprehension of metaphors. The first was the Paraphrase Judgment task, where models were asked to identify the apt paraphrase of the metaphorical sentence. A fill-in-the-blank task was created to crowdsource the apt paraphrases, which were then selected manually. The second task was the Paraphrase Generation Task, where the model was asked to generate a suitable paraphrase for metaphorical sentences.

The results showed that the models faced challenges in identifying apt and inapt paraphrases due to their reliance on lexical similarity, leading them to choose an inapt paraphrase with a similar word to the metaphorical term. They also struggled to differentiate between source and target domains. Additionally, their performance was influenced by text genre, metaphor novelty, and the part of speech (POS) of the metaphorical word.

Baytelman et al. (2024) conducted a study to examine how AI models, mainly ChatGPT and Google Gemini, interpret figurative language and the reliability of their interpretations. Both models were tasked with interpreting metaphors, parables, and other expressions with indirect meanings. Manual and automated experiments were performed using accurately selected data across various fields. The researchers used a Python script to query the ChatGPT API to generate short explanations for 400 English idioms. The experiment demonstrated that while ChatGPT performed well overall, it misinterpreted 2.6% of the idioms. According to Baytelman et al. (2024), this misinterpretation occurred because ChatGPT often focused on literal meanings rather than figurative ones. For instance, the idiom "back seat driver" was interpreted as "someone who always gives instructions and criticism about how a driver should drive from the backseat of the car," narrowing its meaning to the context of driving only.

The intended figurative meaning applies to anyone who provides unwanted advice. Insufficient training of AI models may prevent complete recognition and interpretation of figurative language. The study showed that when figurative expressions are precisely quoted, AI models like ChatGPT process them more effectively than when paraphrased or slightly altered. Providing more context to ChatGPT did not improve interpretations and often caused confusion. Ethical issues may also limit ChatGPT's ability to convey the intended meaning of figurative expressions. For example, the phrase "you cannot teach an old dog new tricks" was interpreted as "it is hard to teach someone resistant to change new skills," without acknowledging that resistance to change is euphemistically associated with old age. Gemini's interpretations were similar to those of ChatGPT, indicating that the issue lies more in information processing than in access to information.

Ichien et al. (2024) conducted a study assessing ChatGPT-4's ability to interpret novel literary metaphors compared to human interpretation. They collected 55 novel literary metaphors from Siberian poems, which underwent a norming process based on quality, metaphoricity, aptness, familiarity, and comprehensibility before being translated into English for reliability. Professional translators translated the metaphors, and the translations were cross-verified for accuracy. The translated versions were then presented to Siberian native speakers competent in English for re-rating based on the same criteria used in the pre-translation norming process. These careful procedures ensured that ChatGPT-4 had not been exposed to these metaphors before and that they were not included in the training database. ChatGPT-4 was asked to generate interpretations of the 55 novel metaphors, and its interpretations were compared to those of 39 undergraduate psychology students at the University of California. In both tasks, the researchers avoided using the term 'metaphor' and instead used the neutral term 'expression.'

The researchers assessed adherence to the Gricean Principle by both ChatGPT-4 and humans by reversing the canonical order of the source and target domains, e.g., "Love is radiance" and "Radiance is love." After gathering interpretations from both, three judges rated their quality on a scale from 0–2, unaware of the AI's involvement. The analysis showed that ChatGPT-4's interpretations were sensible and comparable to human responses, especially with non-canonical metaphors. Both humans and ChatGPT-4 tended to revert to the canonical form, suggesting that ChatGPT-4 has developed pragmatic skills and is sensitive to the Gricean principle. Notably, some of ChatGPT-4's interpretations were rated superior to those of undergraduate students by judges unaware of the former one, and this might be attributed to AI's participation. This contrasts with previous studies, likely due to enhancements in ChatGPT-4 compared to version 3.5.

The literature on AI's ability to interpret figurative language highlights both insights and limitations, emphasizing the need for further research. Earlier systems, such as Hobbs et al.'s ATT-Meta and Narayanan's metaphor-understanding system, laid the groundwork for metaphor interpretation but mainly focused on isolated cases and lacked empirical validation across diverse contexts and languages, raising questions about their relevance to contemporary AI technologies.

Wachowiak and Gromann's (2023) investigation into ChatGPT-3 showed its strengths in predicting source domains but achieved an average accuracy of only 60.22%, with significant shortcomings in processing Spanish. The model's reliance on a limited dataset raises doubts about the robustness of these findings, suggesting they may not accurately reflect broader linguistic patterns, especially in non-English contexts. Additionally, errors such as hallucinations and misidentifications of metaphorical elements indicate ongoing challenges in AI's understanding of metaphors, highlighting a need for deeper linguistic and contextual comprehension.

Table 1 Distribution of the items across the two tests given to Jordanians and Emiratis.

	Test 1 given to Jordanians	Test 2 given to Emiratis
Colloquial metaphorical expressions	20 colloquial Jordanian metaphorical expressions	20 colloquial Emirati metaphorical expressions
Classical Arabic metaphorical expressions	The same 20 Classical Arabic metaphorical expressions	The same 20 Classical Arabic metaphorical expressions
Total	40 metaphorical expressions	40 metaphorical expressions

Similarly, Tong et al.'s (2024) Metaphor Understanding Challenge Dataset (MUNCH) demonstrated that even advanced AI systems like ChatGPT-3.5 and LLaMA struggled with metaphorical expressions, often unable to distinguish appropriate from inappropriate paraphrases. Their focus on lexical similarity reveals a flaw, as it neglects the richness of linguistic context and the cognitive processes supporting human metaphor comprehension. The influence of genre and novelty complicates AI performance, suggesting that current models are limited in adaptability across different linguistic situations.

A comparative study by Baytelman et al. (2024) supported these findings, showing that while ChatGPT and Google Gemini performed reasonably well with idiomatic expressions, their tendency towards literal interpretations highlights a shortcoming in grasping figurative language. Their reliance on direct meanings and context further emphasizes AI's limitations, indicating that despite advancements, these models must evolve to better address the complexities of culturally rich language.

Ichien et al. (2024) evaluated ChatGPT-4 against human interpretations of novel literary metaphors. The study found that ChatGPT-4 produced commendable interpretations, particularly with non-canonical forms, but lacked a robust comparison of AI performance in culturally diverse contexts—an essential aspect for assessing the model's real-world language applicability.

These studies reveal a gap in the assessment of AI's capabilities in interpreting metaphorical language, especially in diverse cultural contexts like Classical Arabic and its colloquial forms. They show that while AI tools have advanced, they still struggle with cultural context and figurative language, which human participants navigate easily due to their experiential knowledge. Additionally, research evaluating AI's efficiency in processing metaphorical language is sparse, mostly focusing on English and other European languages, such as Spanish. No studies have investigated AI's efficiency in processing metaphors in Arabic. Our study aims to address these gaps by conducting a comparative analysis of human participants and four state-of-the-art AI models interpreting metaphors across different Arabic dialects. This exploration is important, as understanding metaphors in culturally rich languages not only improves AI language processing but also helps develop more sophisticated AI tools that align more closely with human cognitive and cultural competencies. The following section describes the methods used to collect and analyze data.

Methods

Sample. Starting with Jordanian participants, there were 29 participants, all native speakers of Jordanian Arabic with working knowledge of Classical Arabic, aged 20–46, twenty of them were university students from the University of Jordan and the other nine participants were employees in different fields and hold different degrees. As far as gender is concerned, the twenty-nine participants were 15 males and 14 females. A purposive sampling technique was employed to adhere to specific inclusion criteria which required the students to be enrolled in any department or school, excluding Arabic literature, linguistics, cultural studies and foreign languages. With respect to the older participants (n = 9), i.e. who are not university students, a convenience sampling technique was used. Participants were recruited directly as they were invited personally, i.e. face-to-face, to participate. The Emirati participants were 21, all native speakers of Emirati Arabic, aged 20–45 (10 females and 11 males). Fifteen were sampled purposively from Mohamed Bin Zayed University for Humanities applying the same criteria above, while 6 older participants were sampled conveniently from the Emirati researcher's acquaintances. Ethical considerations were maintained; informed consent was secured from participants, ensuring confidentiality, anonymity, and the right to withdraw at any time without any consequences (see Appendix 1).

It is important to recognize that choosing participants can introduce some biases. The prevalence of university students may limit the range of educational backgrounds, which could impact how broadly the findings can be applied. Furthermore, while there is a fairly balanced gender representation, it may not fully reflect the different ways people interpret metaphors based on gender. These aspects should be taken into account when we look at the results and what they mean for understanding metaphor comprehension in various cultural contexts.

Materials and data collection. This study employed two distinct tests, each comprising 40 items. Test 1 was administered to Jordanian participants and included 20 metaphorical expressions in Jordanian Arabic and 20 metaphorical expressions in Classical Arabic. Similarly, Test 2 was administered to Emirati participants and contained 20 expressions in Emirati Arabic and 20 expressions in Classical Arabic (see Appendix 1, 2). Table 1 below shows the distribution of the items across the two tests given to the participants.

The colloquial expressions from Jordanian and Emirati cultures were first identified through a pilot study carried out by all four researchers. They observed natural conversations in various informal settings, like casual gatherings and social interactions and wrote down metaphorical expressions as used in context. This observational method enabled the researchers to take detailed notes on a wide range of metaphorical expressions used by native speakers in their everyday speech. To ensure that these expressions were valid and contextually relevant, the researchers used a two-step validation process. First, the research team came together to review the collected expressions, verifying their authenticity within their cultural contexts. This discussion focused on the meanings and usage scenarios of each expression to confirm that they accurately represented colloquial language.

Next, the researchers searched on Facebook, using the expressions as search terms to find examples from actual social media conversations. This analysis offered insights into how these expressions were used in real life, further validating their relevance and usage among speakers. Through this collaborative approach, we aimed to ensure that the selected metaphorical expressions truly reflected the colloquial language of both Jordanian and Emirati participants.

Expressions in Classical Arabic for both tests were sourced from the AlDiwan.net encyclopedia of Arabic poetry. An Arabic literature professor at the University of Jordan reviewed and authenticated these expressions to ensure linguistic accuracy. The expressions across all three Arabic varieties (Jordanian, Emirati, and Classical) encompassed a range of conceptual metaphors, including those based on animals, body parts, container image schemas, spatial image schemas, and other forms of figurative language, examples are:

1. بتحس انه شعبنا فيوزاته ضاربة (Jordanian Arabic JA)
bitħiss ʔinnuh ʃaʕbna: fju:za:tuh dˤa:rbih
Lit. Do you feel like our people's fuses are blown?
'Our people are not thinking clearly'.
2. محمد ذيب ما ينخاف عليه (Emirati Arabic EA)
mħammad ði:b ma: jinxa:f ʕale:h
Lit. Mohammad is a wolf, you don't have to worry about him.
'Mohammed is courageous and resourceful".
3. العَينُ تَذرُفُ وَالفُؤادُ يَذوبُ (Classical Arabic CA)
ʔalʕajnu taðrifu wa lfuʔa:du jaðu:bu
Lit. The eye sheds tears, and the heart melts.
'I am sorrowful and overwhelmed'.

The metaphors in the collected expressions were identified based on MIP Pragglejaz Group (2007) and then the conceptual metaphors were extracted from the metaphorical expressions based on Steen's (2007) steps as follows:

- **Metaphor Identification Procedure (MIP):**

1. Initial Reading: The expression suggests that شعبنا 'our people' are currently experiencing dysfunction or confusion.
2. Identifying Lexical Units: The key metaphorical lexical unit is فيوزاته ضاربة 'fju:za:tuh dˤa:rbih', which translates literally to "their fuses are blown."
3. Literal Meaning: فيوزات 'fuses' refers to safety devices in electrical systems that protect circuitry by breaking the connection when overloaded and ضاربة 'blown' indicates a failure to function properly.
4. Contextual Meaning: In context, this phrase metaphorically suggests that people's ability to think clearly has been compromised, likely due to stress or conflict.
5. Determination of Metaphor: The expression uses "fuses" to symbolize mental clarity and rational thought, which are not physical objects. Thus, it qualifies as a metaphor. The meaning is transferred from the physical realm of machines to the abstract realm of human emotions and thought processes.

- **Extracting the Conceptual Metaphor Using Steen's Steps**

1. Determine if the expression is metaphorical or literal: The phrase فيوزاته ضاربة 'fju:za:tuh dˤa:rbih' is metaphorical, conveying a non-physical meaning regarding mental clarity based on MIP.
2. Identify the underlying propositions related to both conceptual domains: - Source Domain: Electrical systems (fuses blowing). - Target Domain: Human thinking (confusion or lack of clarity). - Proposition: The state of human thought can be compared to the functionality of machines.
3. Form an open comparison between the propositions of the two conceptual domains: - SIM {∃F ∃a [F (HUMAN HEAD)]t [MACHINE (a)]s}, where the human head is perceived as a machine in which a blown fuse corresponds to the failure of clear thinking (see Zibin et al. 2024).
4. Convert the open comparison into a closed one with a formal analogy: - "THE HUMAN HEAD IS A MACHINE SO BLOWING FUSES IS LOSING MENTAL CLARITY." - The act of thinking clearly can be hindered much like a machine that ceases to function properly when its fuses are blown.
5. Establish mappings between the domains: - The state of "blown fuses" correlates with a lack of mental clarity or capability. - The action of "blowing a fuse" maps to the emotional state of being overwhelmed, suggesting that when under pressure, one may "overheat" mentally, similar to a fuse in a machine subjected to excess current.

After identifying metaphors by the researchers, each test, presented on paper, included the 40 expressions with a specific expression underlined in each item. Participants were instructed to provide interpretations solely for the underlined expressions. The term "metaphor" was deliberately omitted to encourage unbiased and natural interpretations (see Ichien et al. 2024). Participants were allotted 25–30 min to complete their respective tests. The tests were administered at the respective campus and in the case of older participants, they were administered at a place of their choosing. Following the test, the researchers conducted short discussion sessions (each lasting 15 min) asking them about the most and least difficult items they encountered and listened to their explanations of their answers. These sessions were conducted to limit any bias in interpreting the participants' answers later. Ethical approval for this research was obtained and all ethical considerations were adhered to (ethical approval document was uploaded).

The selection of ChatGPT-4, ChatGPT 3.5, Google Gemini, and Ask PDF includes top models in natural language processing, recognized for their skills in understanding and interpreting text. ChatGPT-4 and ChatGPT 3.5 were chosen for their effectiveness in language comprehension and metaphor interpretation. Google Gemini was added to examine its innovative methods for interpreting textual data, providing a valuable comparison point. Ask PDF was included for its ability to analyze text from various sources, offering a unique perspective on metaphor extraction. This selection of AI tools enhances our evaluation of the strengths and weaknesses of modern models in processing metaphorical language, which enhances our understanding of AI capabilities in this relatively unexplored area. The four AI tools (ChatGPT-4, ChatGPT 3.5, Google Gemini, and Ask PDF) were also administered both tests, albeit with modifications to the initial administration. The first attempt to provide complete sentences to the AI tools resulted in general descriptions that lacked focus on the underlined expressions and yielded interpretations in English. To rectify this, the tests were re-administered to the AI, with the target expressions clearly highlighted and a specific request for interpretations in Arabic.

Data analysis. Our data analysis was conducted by comparing the interpretation of metaphors by AI and human participants among Jordanian and Emirati Arabic speakers as well as the four AI tools. We employed the Mann–Whitney U test to evaluate differences in accuracy and interpretation between AI tools and human participants from both regions in the contexts of colloquial and classical Arabic. The Mann–Whitney U test is a non-parametric statistical method, which does not assume a normal distribution of the data (MacFarland and Yates 2016). Given that our data comprises performance scores from different groups (AI = 4, Jordanian Arabic = 29, Emirati Arabic = 21), the distribution of scores may not be normal due to variability in the number of participants. This test is specifically designed to compare differences between two independent groups. In our study, we compared the performance of AI tools against two distinct human groups: Jordanian and Emirati speakers, in their understanding of metaphors. The Mann–Whitney U test allowed us to determine whether there are statistically significant differences in interpretation accuracy between these groups (see

Table 2 Mann-Whitney U test result for the differences between AI tools and humans in interpreting metaphors in the two varieties of colloquial Arabic.

Group	N	Mean	Std. Deviation	Accuracy	Mann-Whitney U	Z	Sig. (2-tailed)
AI	4	8.75	3.40	43.8%	1.50	−3.150	0.002
Jordanian	29	17.52	2.18	87.6%			
AI	4	15.50	1.29	77.5%	25.50	−1.232	0.218
Emirati	21	13.81	2.93	69.1%			

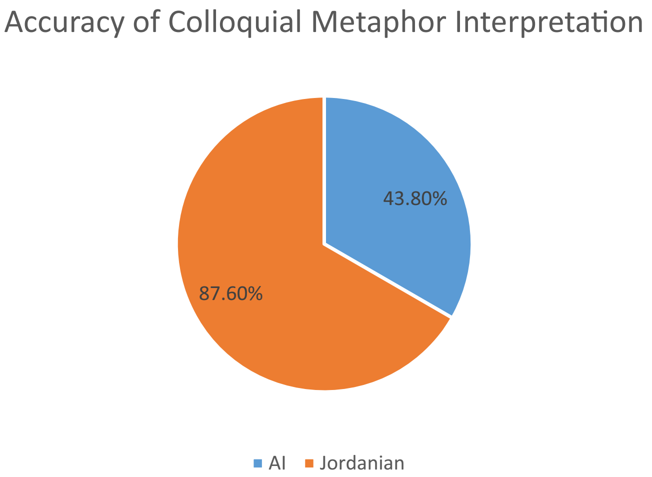

Fig. 1 Accuracy of colloquial metaphor interpretation by AI and Jordanians.

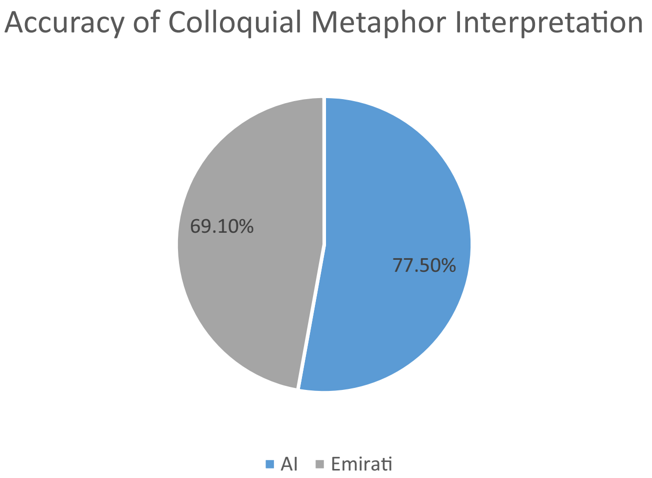

Fig. 2 Accuracy of colloquial metaphor interpretation by AI and Emiratis.

MacFarland and Yates 2016). The results of the data analysis are presented in the following section.

Results

The research questions in this study focus on the effectiveness of AI tools in interpreting metaphors across different varieties of Arabic, as well as how these AI interpretations compare to human interpretations. The analyses conducted using the Mann–Whitney U test provide answers to the research questions (see section "Introduction").

Table 2 presents the Mann–Whitney U test result for the differences between AI tools and humans (both Jordanians and Emiratis) in interpreting metaphors in colloquial Arabic (i.e. Jordanian Arabic and Emirati Arabic), and Figs. 1, 2 provide charts of the accuracy of colloquial metaphor interpretation by AI and Jordanians on the one hand and AI and Emiratis on the other.

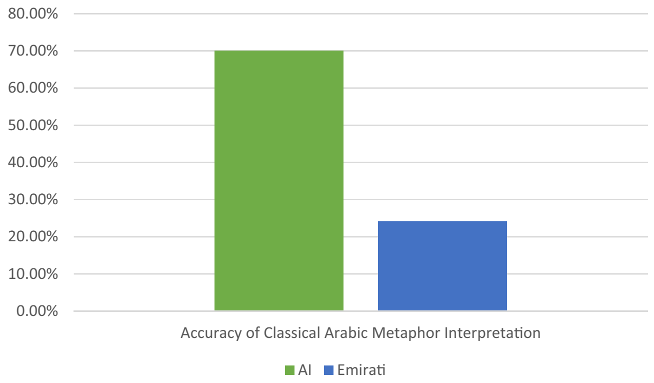

Fig. 3 Accuracy of Classical Arabic metaphor interpretation by AI and Emiratis.

The analysis displayed in Table 2 indicates different levels of success for AI tools in interpreting metaphors across the two Arabic varieties. The Mann–Whitney U test results show a significant difference in performance between AI tools and Jordanian participants, with Jordanians achieving a mean accuracy of 87.6% compared to AI's 43.8% ($Z = -3.150$, $p < 0.05$). This suggests that the AI struggled with colloquial Jordanian metaphors compared to human interpreters. However, no significant difference was observed between AI and Emirati participants ($Z = -1.232$, $p > 0.05$), indicating that the AI's performance was comparable to and slightly higher (mean = 15.50) than that of Emirati participants (mean = 13.81), who had a mean accuracy of 69.1%. The highest accuracy rate was recorded for Google Gemini and ChatGPT-4 while the lowest was recorded for ChatGPT 3.5.

In Classical Arabic (CA), Table 3 presents the Mann–Whitney U test result for the differences between AI tools and humans (both Jordanians and Emiratis) in interpreting metaphors in Classical Arabic.

Table 3 demonstrates that AI tools demonstrated higher accuracy compared to Emirati participants, achieving a mean accuracy of 70.0% versus Emirati's 24.1% ($Z = -2.683$, $p < 0.05$), as shown in Fig. 3.

This indicates that AI was more successful at interpreting classical metaphors than its Emirati counterparts. Conversely, despite the differences in accuracy rates, there was no significant difference between AI and Jordanian performance in Classical Arabic ($Z = -0.166$, $p > 0.05$), revealing that both groups had similar levels of understanding with AI tools having a higher mean of correct answers (mean = 14) than that of Jordanians (mean = 11.93), as shown in Fig. 4:

In addition, there is a statistically significant difference in the interpretations of metaphors in Classical Arabic between Jordanians and Emirati participants in favor of Jordanians ($Z = -4.488$, sig. <0.05), as shown in Fig. 5:

Once again, the highest accuracy rate was recorded for Google Gemini and ChatGPT-4 while the lowest was recorded for

Table 3 Mann-Whitney U test result for the differences between AI and Jordanian and UAE in classical Arabic.

Group	N	Mean	Std. Deviation	Accuracy	Mann-Whitney U	Z	Sig. (2-tailed)
AI	4	14.00	0.82	70.0%	55.00	−0.166	0.868
Jordanian	29	11.93	4.88	59.7%			
AI	4	14.00	0.82	70.0%	6.00	−2.683	0.007
Emirati	21	4.81	4.43	24.1%			
AI	4	14.00	0.82	70.0%	61.00	−1.292	0.196
Jordanian & Emirati	50	8.94	5.85	44.7%			
Jordanian	29	11.93	4.88	59.7%	77.00	−4.488	0.000
Emirati	21	4.81	4.43	24.1%			

ChatGPT 3.5. The results in general provide evidence of the extent to which AI tools can interpret metaphors in both colloquial and classical contexts. While AI's performance in classical Arabic showed promise, its interpretation of colloquial metaphors was significantly less effective compared to human interpreters, particularly among Jordanians. These results are discussed in the following section.

Discussion

The presented results highlight the complexities involved in AI metaphor interpretation concerning human understanding. Even though AI tools exhibit potential in Classical Arabic contexts, their performance in colloquial situations is notably inferior to that of Jordanian participants. This may suggest that cultural and contextual aspects are challenging to AI tools as they may struggle to comprehend them, particularly in conversational language (see Baytelman et al. 2024; Tong et al. 2024). For instance, when asked to interpret the term تسس 'tisss' in the following:

4. . لو عملت قناة يوتيوب فيها محتوى عن مواضيع معينة مع شخص صديق الي بتدعموني ولا هتكونو تسس؟
law ʕmilit qana:t jutju:b fi:ha muħtawah ʕan mawa:dˤi:ʕ muʕajjaneh maʕ ʃaxsˤ sˤadi:q ʔili btidʕamu:ni: willa hatku:nu tisss? (JA)

Lit. If I created a YouTube channel with content on specific topics with a friend of mine, would you support me or would you be tisss?

'Would you be deceitful as a snake?'

GPT-4o mistakenly defined it as sarcasm or mockery, suggesting it meant 'you won't take it seriously.' However, in the context of the expression in (4), the intended meaning was actually 'you are as deceitful as a snake,' referencing the sound of a snake as a metaphor for deceitful behavior. GPT-4o may have difficulty recognizing that تسس 'tisss' represents the sound of a snake and symbolizes deceit for several reasons. First, the training data might not have included enough examples highlighting specific cultural references, leading to a lack of contextual understanding. Additionally, the model's exposure to colloquial Arabic can be limited, making it more dependent on formal interpretations. The ambiguity of language adds complexity, as تسس might not provide sufficient context to trigger the metaphorical connection to deceit. Finally, cultural references associated with this expression may not be universally recognized in the model's training data, causing it to miss the intended meaning that relies on shared cultural knowledge.

Similarly, Ask PDF misinterpreted عظام الرقبة 'neck bones' in:

5. في ناس مصنفة من عظام الرقبة. (JA)
fi: na:s msˤannafeh min ʕðˤa:m ʔirragabeh.

Lit. Some people are thought of as neck bones.

'Some people are very close and important.'

In (5), Ask PDF interpreted the expression as referring to very stubborn and difficult individuals, describing them as الشخصيات عنيدون جداً. This interpretation overlooks the cultural significance of the phrase, which actually implies that such individuals are important, akin to 'the backbone' of a situation. However, without recognizing this cultural meaning, Ask PDF likely missed the intended metaphorical meaning that draws on an understanding of the social and cultural contexts within the language.

In general, several key factors contribute to the disparities in metaphor interpretation scores among Emirati participants, Jordanian participants, and AI tools. Jordanians were shown to have a good grasp of metaphors used in their dialect and in Classical Arabic compared to the Emirati participants. Firstly, this could be attributed to Jordanians' frequent use of these expressions in their daily life conversations (see Mohammed and Ho-Abdullah 2021).

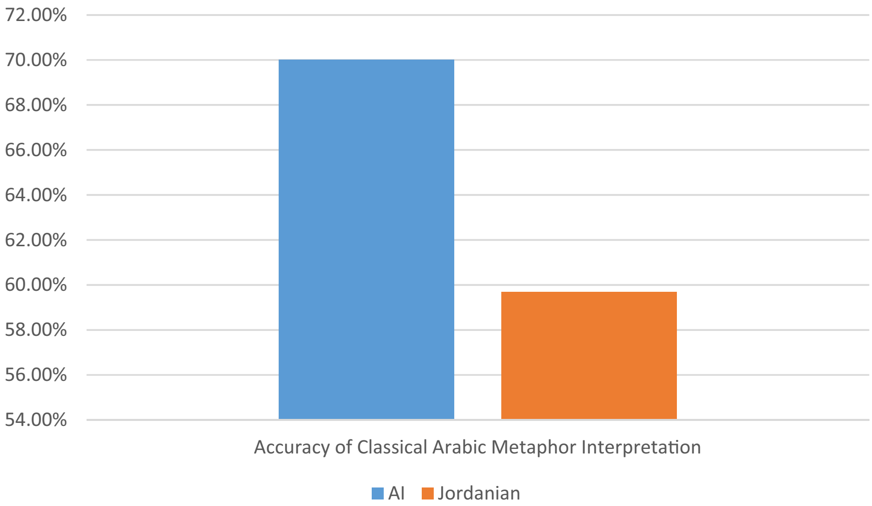

Fig. 4 Accuracy of Classical Arabic metaphor interpretation by AI and Jordanians.

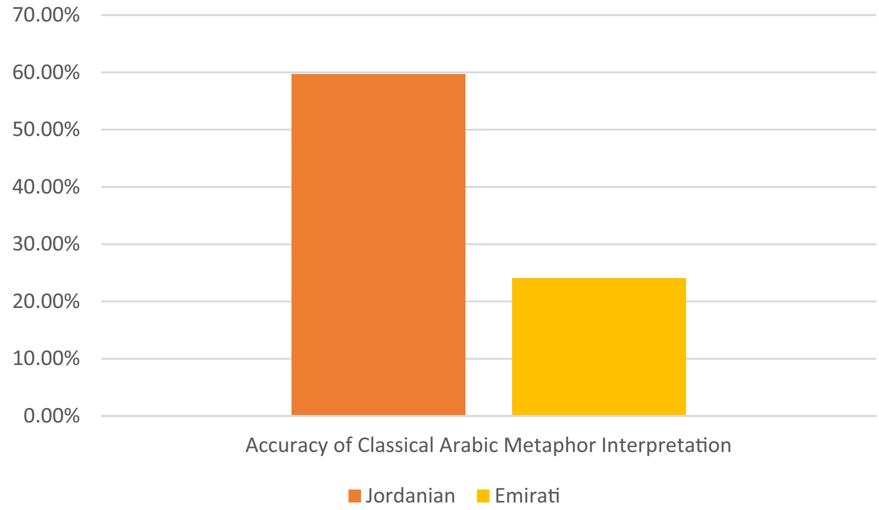

Fig. 5 Accuracy of Classical Arabic metaphor interpretation by Jordanians and Emiratis.

Secondly, it can be suggested Jordan's unique linguistic environment enhances the metaphor interpretation skills observed among Jordanians, particularly in Amman, which serves as a melting pot for various Arabic-speaking communities. As Al-Wer (2007) notes, Amman's diverse demographic landscape features a wide range of Arab cultures, including Palestinians, Syrians, Iraqis, and Chechens, all of which have influenced the Jordanian dialect over time (see also Zibin et al. 2024). This cultural contact could have fostered a blend of metaphors and idiomatic expressions, with the local vernacular increasingly enriched by the influx of diverse linguistic traditions. Consequently, Jordanians encounter a wide array of figurative language in their daily interactions, enhancing their exposure to and familiarity with metaphorical constructs. This linguistic richness potentially allowed Jordanians to develop a heightened sensitivity to the figurativeness of colloquial expressions, enabling them to recognize, interpret, and utilize metaphors more proficiently than their Emirati counterparts, whose dialect may feature a more limited set of idiomatic expressions.

Jordanians demonstrate a strong proficiency in interpreting metaphors in Classical Arabic, a skill that can be attributed to their national education system, which places a significant emphasis on literary criticism, poetry, and the grammar of both Classical and Standard Arabic. Furthermore, students are encouraged to write compositions in Arabic, enhancing their understanding of the language. In contrast, Emirati students may not receive comparable literary training, potentially limiting their sensitivity to metaphorical interpretations. Nonetheless, these observations warrant further research to validate the differences in metaphor interpretation between the two groups.

Finally, in relation to AI tools' performance, AI tools are typically trained on diverse datasets and that usually leads to strong performance in interpreting formal or Classical Arabic expressions. However, this training presents challenges in interpreting colloquial expressions that require an understanding of the cultural context (see Baytelman et al. 2024). This could explain why Emirati participants performed similarly to AI in the interpretation of colloquial expressions. The latter could have had varying levels of exposure to metaphors used in daily conversations which could have contributed to their lower accuracy in metaphor interpretation compared to their Jordanian counterparts. In general, the performance differences between Jordanian and Emirati participants can likely be attributed to a combination of cultural, educational, and linguistic factors. Jordanians, in particular, demonstrate a higher proficiency in interpreting metaphors in both colloquial and classical contexts.

To clarify the differences in metaphor interpretation scores among AI tools, Jordanians, and Emiratis based on the provided expressions, we will employ conceptual metaphor theory as based on main meaning focus (Kövecses 2011) to analyze examples from the test that compare the highest and lowest scoring expressions across the three groups. Note that we are presenting the expressions only without context for word limitation. The entire items are available in Appendices 1 and 2. In addition, note that the explanations provided below were taken from the participants' own reflections on their answers on the tests in the case of humans.

Highest scoring expressions. For Jordanian Arabic participants, the following 2 expressions received the highest scores:

6. عيشته سودا (JA)
ʕe:ʃtuh so:dah
Lit. His life is black.
'His life is miserable.'

The conceptual metaphor used is BAD IS BLACK implying a negative state that involves hardships and misfortune (main meaning focus). The color black generally has negative connotations in JA, making it relatable and easily interpreted. Historically, the color black has often been associated with times of instability and loss, reinforcing its negative connotations. This metaphor may resonate deeply with many Jordanians who are dealing with economic challenges and social pressures. Although black can represent power or elegance in other cultures, its negative associations in Jordanian Arabic make this metaphor especially impactful and relatable.

7. خرفنته (JA)
xarfanatuh
Lit. She made him a sheep.
'She made him docile and submissive.'

Metaphorically, example (7) suggests that he has become submissive to her, implying a loss of autonomy or will (main meaning focus). The reference to a sheep—an animal typically associated with docility and obedience—creates imagery that illustrates HUMAN BEHAVIOR in terms of ANIMAL BEHAVIOR (Goatly 2006). This not only provokes curiosity but also remains relatable through familiar concepts. This metaphor might also reveal deeper anxieties about gender roles and power dynamics, painting a picture where a woman has too much influence over a man.

While it can be seen as funny, it also highlights cultural beliefs about male strength and female manipulation in Jordan.

The following are the two highest scoring items from Classical Arabic.

8. اشتعل الشيب (CA)
ʔiʃtaʕala ʃʃajbu
Lit. Gray hair ignited.
'He aged quickly.'

The conceptual metaphor here compares the rapid spread of gray hair to the rapid spread of fire (main meaning focus). This metaphor suggests that just as fire can spread rapidly and dramatically, the onset of gray hair—representing aging or stress—can also appear suddenly and proliferate over time or through intense experiences. Familiarity with the Quranic verse "وَاشْتَعَلَ الرَّأْسُ شَيْبًا," which translates to ", and my head has become ignited with white hair" in Surah Maryam (Chapter 19) could have helped the participants interpret the metaphor, reinforcing its connotation of rapid aging and the emotional burden that accompanies life's challenges.

9. مهرة عربية وحمار ناهق (CA)
muhratun ʕarabijjatun wa ħima:run na:hiqun
Lit. A purebred mare and a braying donkey.
'She is of high status while he is of low status.'

Again, the conceptual metaphor used here is HUMAN BEHAVIOR IS ANIMAL BEHAVIOR (Goatly 2006; ElShami et al. 2023) contrasting the high status of mares for Arabs with the low status of donkeys reflecting the mappings value and hierarchy (main meaning focus). The example highlights a clear difference in how horses and donkeys are viewed in Arab culture, rooted in deep-seated cultural values. Purebred Arabian mares are highly valued, representing beauty, nobility, and a prestigious lineage. They are often seen as symbols of wealth and power. On the other hand, donkeys tend to have a less favorable reputation, often being seen as stubborn and lacking refinement. Their braying adds to this perception, emphasizing a sense of clumsiness. This contrast serves as a metaphor for wider social hierarchies, underscoring the significance of lineage and status in Arab culture.

The clear conceptual juxtaposition may have improved understanding. In addition, participants mentioned that when interpreting CA expressions. They remembered Arabic lessons where they thought about the subject being compared to the object of comparison as discussed in these lessons and that helped them interpret these expressions.

For Emirati Arabic participants, the following 2 expressions received the highest scores:

10. قلبه أبيض (EA)
galbuh ʔabjadˤ
Lit. His heart is white.
'He is a good-hearted person'.

The example employs color symbolism to convey qualities such as purity, sincerity, and kindness (main meaning focus). In various cultures including Arab culture, white is associated with positive traits such as innocence, tranquility, and goodwill. White is often linked to purity and innocence in many cultures, but in Emirati culture, its meaning runs even deeper. It is closely tied to traditional values and the beauty of the desert landscape. The color white is frequently seen in traditional clothing, where it represents cleanliness and peace. Even more significant is the metaphor of a 'white heart,' which reflects the strong emphasis on generosity, hospitality, and openness that are so important in Emirati culture. Someone described as having a 'white heart' is viewed as warm, kind, and eager to share their blessings with others.

11. طايرة من الفرح (EA)
tˤa:jrah min lfarħah
lit. I am flying from happiness.
'I am extremely happy.'

This example reflects the conceptual metaphor HAPPY IS UP (Lakoff and Johnson 2003). The term "flying" connotes a sense of upward movement or elevation. This spatial orientation is often associated with positive emotions such as happiness, joy, and excitement (main meaning focus). Furthermore, in today's Emirati culture, where travel and global connections are becoming more common, the idea of flying often brings feelings of excitement, adventure, and endless possibilities.

For AI tools, the following JA expressions obtained the highest scores:

12. عيشته سودا (JA)
ʕe:ʃtuh so:dah
Lit. His life is black.
'His life is miserable.'

13. جنن سماي (JA)
dʒannan sama:j
lit. He drove my sky crazy.
'He drove me crazy.'

The following EA expressions obtained the highest scores:

14. قلبه أبيض (EA)
galbuh ʔabjadˤ
lit. His heart is white.
'He is a good-hearted person'.

15. محمد ذيب (EA)
mħammad ði:b
Lit. Mohammad is a wolf.
'Mohammed is courageous and resourceful".

CA expressions:

16. المطبخ السياسي (CA)
ʔilmatˤbax ssija:si:
Lit. The political kitchen.
'The political backstage'.

17. مخضراً جنابه (CA)
muxdˤarran dʒana:buhu
Lit. His side is green.
'He is prosperous and wealthy'.

When asked why these expressions scored so highly, the AI tools explained that they recognize language patterns thanks to the extensive training data they have been exposed to, which helps them spot common metaphors and idioms. However, a closer look shows that the AI actually uses a mix of identifying keywords, mapping semantic relationships, and analyzing contextual cues. AI's knack for picking out metaphors and idioms really depends on the quality and variety of its training data. If the data leans too much towards specific dialects or does not include enough culturally relevant examples, AI might not perform as well. Additionally, while AI can recognize patterns, it often struggles to fully understand the subtleties of context and cultural aspects that are crucial for interpreting metaphors accurately. For instance, AI might have difficulty with metaphors that have multiple meanings or require a deep understanding of the social dynamics at play.

Lowest scoring expressions. For JA participants, the following expressions obtained the lowest scores:

18. فيوزاته ضاربة (JA)
fju:za:tuh dˤa:rbih
Lit. Our people's fuses are blown.
'Our people are not thinking clearly'.

The expression uses a mechanical metaphor to convey that individuals are not thinking clearly. This metaphor may pose interpretive challenges for some, as it depends on an understanding of machinery. While many people know a bit about electricity, only those in technical fields may really understand how fuses work. The metaphor comparing mental clarity to a fuse can feel a bit abstract and might not resonate as well as other, more straightforward expressions. Additionally, there might be simpler ways to convey the idea of ‘not thinking clearly’ in Jordanian Arabic that people find easier to relate to.

19. انشمسنا (JA)
ʔinʃamasna
Lit. We got sunshined.
‘We got exposed and ridiculed.’

The metaphor carries a negative connotation, “sunshined” suggests being thrust into the spotlight in an uncomfortable or vulnerable manner (main meaning focus), where one’s actions are subject to scrutiny and laughter. This expression lacks direct imagery or an obvious conceptual connection, making it harder for interpreters to grasp its meaning. This metaphor might also touch on fears of public shaming, which might be a major cultural worry in Jordan. Additionally, the expression itself might be fairly new or not widely used, making it harder for people to grasp its meaning.

Regarding CA, the following received the lowest scores:

20. ملح الأرض (CA)
milħu lʔardˤ
Lit. The salt of the earth.
‘They are essential’.

While the metaphor usually suggests something is essential, it does not specify what kind of essentialness is being conveyed. It could mean moral, practical, or social importance, which can be confusing. Additionally, the expression might be interpreted in different ways, possibly suggesting that these individuals are too conservative or resistant to change. This lack of clear meaning likely played a part in its low score.

21. لهم في لحوم العالمين مخالب (CA)
lahum fi: luħu:m lʕa:lami:n maxa:lib
Lit. They have claws in the flesh of the worlds.
‘They are powerful/ aggressive.’

This metaphor suggests that this group is deeply involved in and significantly impacts various societies. Their “claws” symbolize aggression, power, and a predatory nature, indicating that they exert control or dominance, potentially in harmful or exploitative ways (main meaning focus). The complex imagery may be interpreted in multiple ways, resulting in ambiguity and confusion.

For EA participants, the following obtained the lowest scores:

22. تكهرب (EA)
tkahrab
Lit. He got electrocuted.
‘He was shocked.’

The term تكهرب functions as a metaphor for a sudden emotional or psychological shock, typically signifying a state of surprise, excitement, or overwhelming realization (main meaning focus). The metaphor may not hold a clear, tangible meaning in a metaphorical context, making it difficult to interpret. The connection between electrocution and emotional shock may also be too indirect or abstract for some speakers.

23. النية مطية (EA)
ʔinnijjah matˤijjah
Lit. Intension is a mount.
‘One’s intention guides their actions.’

The expression النية مطية serves as a metaphor that emphasizes the significance of one’s intentions in guiding actions and outcomes. The term “mount” suggests that intentions can propel an individual forward, akin to how a vehicle facilitates travel. The metaphor may not readily convey meaning, resulting in lower comprehension scores. The word ‘مطية‘ (mount) may also be considered somewhat formal or outdated, and its use in everyday Emirati Arabic may be limited.

Concerning CA expressions, the following obtained the lowest score:

24. يداً نديه صفراء (CA)
jadan nadijjah sˤafra:ʔ
lit. Moist yellow hands.
‘Hands stained yellow from poverty’.

The expression can be interpreted as ‘hands stained yellow from poverty.’ The ‘yellow’ hue symbolizes not only the physical toll of their circumstances but also the economic and social challenges they face. This imagery was difficult to interpret for EA participants. In Arab culture, the color yellow can sometimes evoke negative feelings, e.g. sickness or envy. Moisture can also be linked to poverty, which might suggest uncleanliness or illness. This imagery is quite distinct and might not immediately resonate with the idea of poverty. These elements likely played a role in its lower score.

Concerning AI tools expressions, the following obtained the lowest scores:

25. عينك ما تشوف الا النور (EA)
ʕejnak ma: tʃu:f ʔilla nnu:r
Lit. Your eyes say nothing but light.
‘Someone disappeared.’

26. ضوء مصباح (CA)
dˤu:ʕu misˤba:ħ
lit. Light of a lamp
‘My beloved has a bright smile’.

27. عينه ضيقة (JA)
ʕejnuh dajjʔa
Lit. His eye is narrow.
‘He is envious of people’.

AI tools answered that the combination of ambiguity, cultural references, non-literal meanings, and emotional aspects can make these expressions difficult for AI to interpret accurately. A closer look shows that AI’s reliance on pattern recognition can actually be a significant drawback when it comes to less common expressions. Although AI excels at spotting patterns in large datasets, it often has a tough time with metaphors that stray from the usual patterns or depend on subtle context clues. Plus, having world knowledge is important to interpreting these metaphors. AI might not have the background information needed to grasp the cultural context, historical references, or social implications that are vital for a proper understanding. For instance, to get the meaning of ‘yellow hands,’ one needs to know about the potential historical links between certain types of work and skin discoloration. Lastly, AI’s ability to deal with figurative language is somewhat limited. Although it can recognize some common metaphors and idioms, it frequently struggles with more complex or unconventional expressions that require a deeper grasp of language and culture.

In summary, the application of the theory demonstrates that the effectiveness of metaphors among AI tools, Jordanians, and Emiratis depends on their clarity, cultural references, and connections to shared experiences within the language (see Alazazmeh and Zibin 2023). Higher-scoring expressions employ concrete imagery and cultural familiarity, enhancing accessibility and comprehension. Conversely, lower-scoring metaphors frequently present challenges due to their complexity, abstraction, or insufficient cultural relevance. Consequently, understanding metaphorical language in Arabic—and its interpretation by both AI and human speakers—depends on these foundational conceptual frameworks.

Compared with previous studies, the current one found that Google Gemini and ChatGPT-4 exhibited high accuracy rates in interpreting metaphors, consistent with previous research that highlights advancements in AI models, particularly ChatGPT-4’s enhanced ability to grasp complex metaphorical language. In contrast, ChatGPT-3.5 showed the lowest accuracy in metaphor interpretation, reflecting findings by Wachowiak and Gromann (2023), which noted its difficulty with idiomatic expressions and context, resulting in only 60.22% accuracy. Moreover, challenges in recognizing culturally significant metaphors and phrases were observed, particularly in colloquial Arabic dialects, similar to Tong et al. (2024). These results suggest that metaphor interpretation in AI requires not only strong linguistic capabilities but also an understanding of cultural context, as emphasized by Ichien et al. (2024).

Conclusion

This study examined the effectiveness of AI tools in interpreting metaphors within Jordanian and Emirati Arabic dialects as well as Classical Arabic and compared their performance to that of human respondents. The analysis revealed differences in accuracy among the AI tools; Google Gemini and ChatGPT-4 outperformed ChatGPT-3.5, which struggled with metaphor interpretation and had limitations in distinguishing between metaphorical and literal meanings. These findings highlight the importance of advancements in AI models for better language processing. Additionally, the study emphasized the critical role of cultural context and familiarity, as human participants, particularly Jordanians, excelled in interpreting colloquial metaphors due to their deep cultural understanding. This highlights that while AI can analyze extensive and formal data, it often lacks the cultural insight necessary for accurate comprehension of figurative language.

The paper contributes to the discourse on AI in language interpretation, pointing out both the potential and challenges of AI in processing culturally rich and context-driven language. Future studies should focus on enhancing AI models through cultural contextualization and improved training methodologies to better support language understanding and communication. That is, these studies should emphasize the enhancement of AI models through the integration of cultural context and improved training methods and the innovative incorporation of cultural references into AI processing algorithms. Collaborations between linguists and AI developers could lead to advancements in creating AI systems that better comprehend figurative language across diverse cultural contexts. Additionally, conducting comparative studies on various regional dialects and languages could reveal the complexities of cultural interpretation and guide the development of more sophisticated AI tools.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 4 December 2024; Accepted: 9 June 2025;
Published online: 01 July 2025

Note

[1] Classical Arabic (CA) was utilized in literary texts from the pre-Islamic and early Islamic periods. It is characterized by its complex grammar and usage rules, which can present challenges for modern speakers. This form of Arabic is primarily employed in religious contexts, including Islamic scholarship, Quranic recitation, and traditional poetry. Modern Standard Arabic (MSA) is a standardized variant that has evolved from Classical Arabic (van Putten 2020). It functions as the lingua franca across the Arab world. MSA simplifies some of the complexities of Classical Arabic while incorporating contemporary vocabulary to accommodate modern life. It is utilized in formal settings such as news broadcasts, literature, official documents, and educational environments (van Putten 2020).

References

Alazazmeh HM, Zibin A (2023) The conceptualization of anger through metaphors, metonymies and metaphtonymies in Jordanian Arabic and English: A contrastive study. Cogn Semant 8(3):409–446. https://doi.org/10.1163/23526416-bja10037

Al-Wer E (2007) The formation of the dialect of Amman: From chaos to order. In Miller, C, Al-Wer, E, Caubet, D, & Watson, JC (Eds.) *Arabic in the City* (pp. 69–90). London: Routledge. https://doi.org/10.4324/9780203933367

Barnden JA (2001) Uncertain reasoning about agents' beliefs and reasoning. Artif Intell Law 9(2):115–152. https://doi.org/10.1023/A:1017993913369

Barnden JA, Helmreich S, Iverson E, Stein GC (1994) An integrated implementation of simulative, uncertain and metaphorical reasoning about mental states. In J Doyle, E Sandewall, and P Torasso (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference* (pp. 27-38). San Mateo, CA: Morgan Kaufmann

Baytelman Y, Baytelman M, Potsepaiev V (2024) Testing AI ability to process figurative language. Sci works Donetsk Natl Tech Univ Probl modeling Des Autom 19(1):89–95. https://doi.org/10.31474/2074-7888-2024-1-19-89-95

Boden MA (2016) Ai: Its Nature and Future. Oxford University Press UK

Clausner TC, Croft WB (1997) Productivity and schematicity in metaphors. Cogn Sci 21(3):247–282. https://doi.org/10.1207/s15516709cog2103_1

Di Biagio, G (2022). *Disruptive metaphors: how to enhance metaphor creation and augment the* innovation *of meaning process with GenAI*. Unpublished MA thesis, Politecnico, Milano

ElShami THS, Shuaibi JA, Zibin A (2023) The function of metaphor modality in memes on Jordanian Facebook pages. SAGE Open 13(1):1–24. https://doi.org/10.1177/21582440231154848

Fass D (1997) Processing Metonymy and Metaphor. Bloomsbury Academic

Gargett A, Barnden J (2013) Gen-meta: Generating metaphors using a combination of AI reasoning and corpus-based modeling of formulaic expressions. In *2013 Conference on Technologies and Applications of Artificial Intelligence* (pp. 103–108). IEEE

Gargett A, Mille S, Barnden J (2015) Deep generation of metaphors. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 336-343). IEEE

Goatly A (2006) Humans, animals, and metaphors. Soc Anim 14(1):15–37. https://doi.org/10.1163/156853006776137131

Group P (2007) MIP: A method for identifying metaphorically used words in discourse. Metaphor Symb 22(1):1–39. https://doi.org/10.1080/10926480709336752

Hobbs JR, Stickel ME, Appelt DE, Martin P (1993) Interpretation as abduction. Artif Intell 63(1–2):69–142. https://doi.org/10.1016/0004-3702(93)90015-4

Ichien N, Stamenković D, Holyoak KJ (2024) Large language model displays emergent ability to interpret novel literary metaphors. Metaphor Symb 39(4):296–309. https://doi.org/10.1080/10926488.2024.2380348

Kövecses Z (2003) Metaphor and emotion: Language, culture, and body in human feeling. Cambridge University Press, Cambridge

Kövecses Z (2008) Conceptual metaphor theory: Some criticisms and alternative proposals. Annu Rev Cogn Linguist 6:168–184. https://doi.org/10.1075/arcl.6.08kov

Kövecses Z (2011) Methodological issues in conceptual metaphor theory. In Handl S, Hans-Jörg S (eds.) *Windows to the Mind: Metaphor, Metonymy and Conceptual Blending* (pp. 23–40). Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110238198.23

Lakoff G, Johnson M (2003) Metaphors we live by. Chicago: University of Chicago Press

Langacker Ronald W (1987) *Foundations of cognitive grammar: Volume I: Theoretical prerequisites* (Vol. 1). Stanford: Stanford University Press

MacFarland TW, Yates JM (2016) *Introduction to nonparametric statistics for the biological sciences using R* (pp. 103–132). Cham: Springer

Martin JH (1990) A computational model of metaphor interpretation. San Diego, CA

Mohammed TMQ, Ho-Abdullah I (2021) Universality and language specificity: Evidence from Arab and English proverbs. Int J Comp Lit Translation Stud 9(1):24–30

Narayanan S (1997) *Knowledge-based action representations for metaphor and aspect (KARMA)* (Doctoral dissertation, PhD thesis, Computer Science Division, EECS Department, University of California at Berkeley)

Narayanan S (1999) Moving right along: A computational model of metaphoric reasoning about events. In Proceedings of the National Conference on Artificial Intelligence (AAAI '99) (pp. 121–129). Orlando, FL: AAAI Press

Steen G (2007) Finding metaphor in discourse: Pragglejaz and beyond. Cult, Leng y Representación/Cult, Lang Representation 5:9–25

Tong X, Choenni R, Lewis M, Shutova E (2024) Metaphor understanding challenge dataset for LLMs. *arXiv preprint arXiv:2403.11810*. https://doi.org/10.48550/arXiv.2403.11810

van Putten M (2020) Classical and modern standard Arabic. In Lucas, C, & Manfredi, S (eds.) *Arabic and contact-induced change* (pp. 65–82). Language Science Press

Veale T (1998) 'Just-in-Time' Analogical Reasoning: A Progressive-Deepening Model of Structure-Mapping. In *the proceedings of ECAI'98, the 13th European Conference on Artificial Intelligence*, Brighton, UK (pp. 93–97)

Wachowiak L, Gromann D (2023) Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1018-1032)

Zibin A, Altakhaineh ARM (2023) A blending analysis of metaphors and metonymies used to depict the deal of the century by Jordanian cartoonists. Lang Cognit 15(2):377–404. https://doi.org/10.1017/langcog.2023.1

Zibin A, Daoud S, Mitib Altakhaineh AR (2024) Indexical meanings of the realization of/s^{ʕ}/ص as [s] س in spoken and written Jordanian Arabic: a language change in progress? Folia Linguistica 58(2):267–290. https://doi.org/10.1515/flin-2024-2003

Zibin A, Altakhaineh ARM, Musmar O (2024) HEAD metonymies and metaphors in Jordanian and Tunisian Arabic: An extended conceptual metaphor theory perspective. Lang Cognit 16(4):2009–2031. https://doi.org/10.1017/langcog.2024.31

Author contributions

Conceptualization, AZ, HA, HY and NB; methodology, AZ, HA, HY and NB; Statistical analysis, AZ and NB; validation, AZ, HA, HY; formal analysis, AZ; investigation, AZ, HA, HY and NB; resources, AZ; data curation, NB, HA and HY; writing—original draft preparation, AZ, HA, HY and NB; writing—review and editing, AZ, HA, HY and NB; supervision, AZ; project administration, AZ.

Competing interests

The authors declare no competing interests.

AI statement

As the authors are not native speakers of English, Google Gemini was employed to assist with proofreading and grammar (accessed 4th Nov 2024; 17th April 2025). The authors carefully reviewed all suggested edits and take full responsibility for the accuracy and integrity of the content presented.

Ethical approval

This study was conducted in accordance with the ethical guidelines of the University of Jordan. The Ethics Committee at the Department of English Language and Literature at The University of Jordan convened on November 29, 2024, and agreed to the study's protocol. Official ethical approval was granted on December 30, 2024, with reference number (1/30/12/2024), following standard bureaucratic procedures.Informed consent Prior to the issuance of the official ethical approval document, informed consent was obtained from all participants on November 30, 2024, through a written format. This process was undertaken with the understanding that the Ethics Committee had already convened on November 29, 2024, and indicated their agreement to the study's protocol. Participants included students at the University of Jordan and Mohammed Bin Zayed University for Humanities, as well as conveniently sampled participants from Jordanian and Emirati backgrounds. The informed consent document was presented to participants in written format to ensure clarity and understanding. Each participant was given the opportunity to ask questions and receive comprehensive answers before providing their signature to signify informed consent. A copy of the signed informed consent form is available in the appendix for reference and verification.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-05282-0.

Correspondence and requests for materials should be addressed to Aseel Zibin.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.