Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Specific Linguistic Questions that Artificial Intelligence (AI) Cannot Answer Accurately: Implications for Digital Didactics

Reima Al-Jarf
Full Professor of English and Translation Studies, Riyadh, Saudi Arabia
Corresponding Author: Reima Al-Jarf**, E-mail:** reima.al.jarf@gmail.com

| ABSTRACT
This study aimed to examine the kinds of specific linguistic questions that AI systems cannot answer accurately, drawing on daily scholarly interactions with Copilot (MC), DeepSeek (DS), Google Translate (GT), Gemini, Monica, rather than artificial test prompts. It also seeks to understand why AI makes mistakes and whether the nature of user questions influences AI performance. In addition, it aims to analyze the nature of these failures through a linguistic and a pedagogical lens to offer recommendations for students and researchers on how to critically evaluate AI output in language and translation tasks. For those purposes, a sample of 50 specific linguistic and translation questions that the author asked to MC, DS, GT, Gemini & Monica between 2023 and late 2025 was collected and anlayzed. The questions were classified into the following categories: phonology, transcription, morphology, lexical questions, pragmatics and culture, explanation or translation of Arabic grammatical terms, books AI cannot fully identify, telling a story in Arabic classical literature, converting a hand-written text to a typed text, translation of technical terms, metaphorical expressions and metonyms, and bibliographic, and scholarly workflow issues. Results of the 5 AI responses to the 50 linguistic and translation questions and tasks revealed recurrent AI shortcomings across diverse domains: phonology and transcription errors, morphological and lexical inaccuracies, pragmatic and cultural misinterpretations, and faulty explanation of Arabic grammatical terms. AI systems failed to identify certain books, struggled with classical storytelling in Arabic classical literature, and could not convert handwritten to a typed text. Technical terminology, metaphorical expressions, and metonyms were often mistranslated, while bibliographic and scholarly workflow tasks showed fabricated references and gaps in organizing references. Collectively, these errors underscore AI's tendency toward surface-level processing at the expense of linguistic depth and cultural fidelity. Findings of this study suggest directions for refining AI translation models and integrating critical AI literacy into language education. This study also contributes to ongoing debates on the limitations of AI and the use of AI in linguistics by highlighting the pedagogical risks of uncritical reliance on AI output. Despite these small flaws and issues, AI is capable of performing unlimited linguistic tasks with lightning speed and impressive content that humans are incapable of.

| KEYWORDS
Artificial Intelligence (AI), AI linguistic limitations, AI translation limitations, AI responses to author initiated questions, AI inaccuracies, causes of AI errors, improving AI responses, translation didactics, studetnts' linguistic skills, students' searching skills.

| ARTICLE INFORMATION
ACCEPTED: 29 November 2025 **PUBLISHED:** 13 December 2025 **DOI:** 10.32996/fcsai.2025.4.4.4

1. Introduction

Nowadays, students and researchers use AI for a variety of tasks in language and translation. For example, AI is good at brainstorming ideas, clarifying tricky concepts, and tasks like recapping meetings and helping research a topic. It can write emails, refine work, generate study plans, quizzes, and practice questions to test one's knowledge, create summaries, get an overview of reports or lecture notes, including key points and takeaways, translate text and checks grammar, and help students practice pronunciation. It can compose a song or story, explore different styles, envision their work across a range of approaches, visual genres, and copy formats (Al-Jarf, 2025l; Al-Jarf, 2024a).

Despite this apparent fluency, AI language models often lack genuine comprehension and common sense, making their output fundamentally unreliable for high-stakes applications.

A review of research studies conducted between 2023 and 2025 revealed numerous limitations of large language models (LLMs) in linguistics and translation such as weaknesses in contextual understanding, pragmatic, lack of cognitive grounding, inaccuracies, gaps between formal and functional competence, focus on narrow technical tasks, not deep linguistic understanding, failure at disambiguation, metalinguistic intuition and dialect handling (Önem, 2025; Uchida, 2024; Mahowald et al., 2024; Vajjala, 2024; Langacker, 2024);

Another group of studies revealed cultural and sociolinguistic gaps (Al Yahya, Al Muoaeweed, & Bilal, 2025; Alaqloobi et al., 2024).

Additionally, numerous studies detected bias, ethical and data limitations as (Vatsadze, 2025; Ahmad, 2024; Tai et al. 2023; Alsharekh, Talavage, & McDonald, 2024).

Technical and linguistic constraints, including failures in long-memory tasks, have also been revealed (Krasnyuk et al., 2023; Noor et al. 2024; Mahowald, 2023; Šprogar, 2024; Cuskley, Woods, & Flaherty, 2024).

Further studies revealed lower quality in Arabic/Chinese compared to English, and the marginalization of endangered, minority and low-resource languages (Jeon et al., 2025; Lynch 2025; Ali, Bhatti, & Abbas, 2025; Kshetri, 2024).

Most importantly, weaknesses in the use of language elements and skills such as grammar/semantic errors, limitations in peer-reviewed and academic writing, generative linguistics, focus on surface syntax, not deep semantics, failure in long-term memory and deep reasoning tasks, poor handling of complex linguistic puzzles and inability to handle nonsense sentences (Baziyad, Kamel, & Rabie, 2023; Shormani, 2024; Jankowska, Jankowski, & Szymański, 2024; Golan et al., 2025; Wang & Usher, 2024).

Furthermore, pedagogical risks, language Learning limitations, loss of learner creativity, problems in acquisition and meaning representation, failure to support critical thinking and authentic communication, risk of mechanistic personalization in learning and struggling with aphasia speech (Ansari & Ansari, 2025; Gerlich, 2025; Jeon et al., 2025; Révész, Suzuki, & Jung, 2025; Zeng et al., 2024).

Most of the above studies relied on experimental outputs from AI systems such as Copilot, DeepSeek, Gemini, ChatGPT, GPT-4, and Monica, and often compared them with human performance or linguistic corpora. Rather than theorizing, researchers posed specific linguistic tasks - grammar analysis, translation, idiom interpretation, and academic writing - and then evaluated AI responses for accuracy, fluency, cultural sensitivity or creativity (Noor et al., 2024; Mahowald, 2023; Baziyad et al., 2023). Other studies compared AI outputs with human translations or writings to identify differences in contextual accuracy and creativity (Jankowska et al., 2024; Alsharekh et al., 2024). Some examined performance in endangered and minority languages by comparing AI with native speakers (Kshetri, 2024; Ali, Bhatti & Abbas, 2025). Corpus-based analyses studies tested vocabulary diversity, syntactic structures, and pragmatic interpretation (Langacker, 2024; Uchida, 2024). In addition, mixed-methods approaches incorporated questionnaires, surveys, and classroom observations to measure how students and teachers perceived AI's role in learning (Révész et al., 2025; Gerlich, 2025). Finally, some studies relied on secondary sources such as content analyses, bibliometric reviews, and policy documents to identify broader trends and ethical concerns (Ansari & Ansari, 2025; Vatsadze, 2025; Alaqloobi et al., 2024).

Most prior studies reviewed above have documented AI's linguistic limitations by relying on controlled experiments, artificial prompts, or corpus analysis. They have not examined how AI systems respond to authentic, unpredictable linguistic questions posed by students and researchers in real world practice. This gap leaves how AI performs when confronted with nuanced, culturally embedded, or technically demanding queries unexplored.

To address these gaps, the current study aims to examine the kinds of specific linguistic and translation questions that AI systems cannot answer accurately, drawing on daily scholarly practice and interactions with Copilot, DeepSeek, Gemini, Monica, and Google Translate rather than artificial test prompts. It also seeks to understand why AI makes mistakes and whether the nature of user questions influences AI performance. In addition, this study aims to Analyze the nature of these failures through a linguistic lens (phonology, grammar, semantics, pragmatics,) and a pedagogical lens (implications for teaching, learning, and digital didactics). It will offer recommendations for students and researchers on how to critically evaluate AI output in language and translation tasks.

This study is not based on pre-determined questions. The questions were asked between 2023-2025, during the author's daily interaction with the five AI systems. Since some questions were asked to one, two or three AI systems, the aim of this study is not to compare the different AI systems, but rather to observe how AI responds in authentic scholarly contexts.

By analyzing faulty responses from multiple AI systems through a linguistic and pedagogical lens, this study provides a field-based testimony to their performance - something previous research has not achieved, as it often remains at the level of general analysis or focuses on a single aspect. It highlights AI's limitations and their implications for translation, language learning, teaching, translation and critical AI literacy. Unlike prior studies that rely on theoretical analysis or controlled experiments, this research draws upon direct, real-world interactions with multiple AI systems. By posing authentic questions that emerge from daily scholarly practice, it documents where AI succeeds and where it fails, offering insights and recommendations for students, instructors and researchers. In doing so, it connects linguistic inaccuracies directly to pedagogical risks, underscoring the need for educators and learners to approach AI critically and to integrate AI literacy into digital didactics.

Furthermore, this study is an addition to a series of studies by the author on the use of AI in translation and education such as: educational polysemes in AI translation of Arabic research articles (Al Jarf, 2025a); encrypted Arabic on Facebook and YouTube (Al Jarf, 2025d); "sleep" terms (Al Jarf, 2025o); Gaza–Israel war terminology (Al Jarf, 2025b); grammatical terms used metaphorically (Al Jarf, 2025h); expressions of impossibility (Al Jarf, 2025m); zero expressions (Al Jarf, 2025p); human vs AI translation of chemical compound names (Al Jarf, 2025i); Arabic *abu* brand names (Al Jarf, 2025e); denotative and metonymic abu- and umm- animal and plant folk names (Al Jarf, 2025g); folk medical terms with *om* and *abu* (Al Jarf, 2025n); medical terms (Al Jarf, 2024b); technical terms (Al Jarf, 2021; Al Jarf, 2016a); electronic translation between Arabic and European languages (Al-Jarf, 2012); Arabic transliteration of borrowed English nouns with /g/ (Al Jarf, 2025c); pronunciation errors in Arabic YouTube videos (Al Jarf, 2025f; Al Jarf, 2025j; Al Jarf, 2025k); editors' perspectives on the publication of AI-generated research articles (Al Jarf, 2025l); Arab instructors' views on AI-generated student assignments (Al Jarf, 2024a. Together, these studies illustrate recurring weaknesses in AI's handling of linguistic, cultural, and scholarly tasks, reinforcing the diagnostic framework adopted in this paper.

2. Methodology

2.1 Data Collection

A sample consists of 50 specific linguistic and translation questions that the author asked to Microsoft Copilot (MC), DeepSeek (DS), Google Translate (GT), Gemini & Monica between 2023 and late 2025. Questions about health, travel, world events, cooking, art and other non-linguistic domains or questions that were correctly and accurately answered, were excluded. The questions herein are real scholarly interactions rather than contrived experiments, spontaneously asked to MC, DS, GT, Gemini and Monica while the author was working on her research or exploring AI capacity to respond to general linguistic issues. Some questions were asked to a single AI, others to two three AI systems. The set of questions included in the current study is not exhaustive but is representative of authentic scholarly needs in linguistics and translation.

The questions were not pre-determined. When the author first started to use AI, it was not her intention to compile the questions and conduct a study. The author's focus, then, was on how AI performs in translating certain types of terms and metaphorical expressions. The idea of compiling the linguistic and translation questions emerged later after extensive engagement with AI systems, and the publication of 20 research articles on the use of AI in translation and education. Each question was posed directly to an AI system in real-time interaction. Responses were archived verbatim to preserve accuracy. The same question was sometimes asked across systems, not for comparative benchmarking but to observe performance in authentic contexts.

2.2 Data Analysis

This study adopts an ecological approach by analyzing authentic, spontaneously posed questions rather than pre-designed prompts, thereby reflecting real scholarly practice. This approach was followed in the current study. No human subjects were involved in the study, so ethical approval was not required.

The questions were classified into the following categories: phonology, transcription, morphology, lexical questions, pragmatics and culture, explanation or translation of Arabic grammatical terms, books AI cannot fully identify, telling a story in Arabic classical literature, converting a hand-written document to a typed document, translation of technical terms, metaphorical expressions and metonyms, and bibliographic, and scholarly workflow.

Each answer was evaluated in terms of grammatical correctness, semantic fidelity, pragmatic adequacy, phonological precision and pedagogical relevance, usability in teaching and learning contexts, risks of misinformation, and potential erosion of creativity or critical thinking. Answers were coded as accurate, partially inaccurate or inaccurate. Percentages were calculated for correct and incorrect responses within the same question. Examples were given to illustrate a particular sound, correctly translated words, terms and metaphorical expressions in each domain. In addition, weaknesses were described verbally and illustrated by examples.

AI responses were coded by the author based on linguistic and pedagogical criteria, with examples provided to illustrate key findings. While this single-coder approach represents a limitation, it is mitigated by the author's 35 years of expertise in linguistics and translation 300 publications (including 20 publications on the use of AI and translation and education), 700 conference presentations in 75 countries mitigate this limitation.

Moreover, results of the current study may not be generalize beyond the specific AI systems (MC, DS, GT, Gemini, and Monica) tested and beyond the specific answers given herein as these answers reflect the AI systems' calibre at a specific point of time especially since some questions were asked in 2023, others in late 2025. Asking these questions, even to the same AI system now, may yield different answers due to the daily improvements in the performance of all AI systems, and their training data, especially since the AI systems under study are learning systems.

3. Results

3.1 Questions that AI does not answer accurately

1) ***Phonological questions***

- **Give me English personal names with g pronounced /g/ not / dʒ/. G can be at the beginning, middle or end**. AI gave 18 names, 78% were correct (Gary, Grant, Gregory, Gwendolyn, August, Douglas, Logan, Megan, Margaret, Craig, Greg, Young, King, Sterling) and 22% were faulty (George Reginald, Roger, Virginia with a j).

- **Give me city and country names containing g pronounced /g/:** AI gave 35 (78%) correct names (Gabon, The Gambia, Ghana, Greece, Grenada, Guatemala, Guinea, Guyana, Congo, Hungary, Luxembourg, Portugal, Singapore, Uganda, United Kingdom, Giza, Guayaquil, Guadalajara, Glasgow, Gold Coast, Guatemala, Gaziantep, Gdynia, Grand Rapids, Gaborone, Graz, Groningen, Portugal, Birmingham, Prague, Bogota, Lagos, Glasgow, Hamburg, Singapore and 10 (22%) faulty names with g pronounced /j/ (Georgia, Germany, Argentina, Belgium, Egypt, Nigeria, Guangzhou, Genoa, Georgetown, Geneva).

- **Give me medical terms with double L:** Gemini gave 18 words, 78% are correct (Tonsillitis, Lell's sign, Cellulitis, Capillaries, Medulla, Pallor, Bullous, Allergy, Allograft, Allele, Papillary, Fallopian tubes, Collagen, Pupillary) and 22% were faulty (Palsy, Cholesterol, Myelin sheath, Glomeruli).

- **Give me 25 words with a silent d in the middle**: AI gave 8 (32%) correct words (Wednesday, Handsome, Sandwich, Handkerchief, Grandsons, Grandfather, Grandmother, Windmill), 3 (12%) faulty words with d in the middle(Brandish, Kindergarten, Granddaughter) and 14 (56%) faulty words ending in dge (Pledge /plɛdʒ/, Hedge /hɛdʒ/, Edge /ɛdʒ/, Bridge /brɪdʒ/, Fudge /fʌdʒ/, Lodge /lɑːdʒ/, Badge /bædʒ/, Grudge /grʌdʒ/, Knowledge /ˈnɑːlɪdʒ/, Nudge /nʌdʒ/, Ridge /rɪdʒ/, Sledge /slɛdʒ/, Budget /ˈbʌdʒɪt/, Adjective /ˈædʒɪktɪv/)

- **Give me 20 words with a silent s:** AI gave 15% correct words (Island, Aisle, Isle), 5% faulty words with no s at all (Patio), and 80% French words used in English with silent s mostly in word final position (Puisne, Viscount, Debris, Bourgeois, Corps, Chassis, Arkansas, Illinois, Rendezvous, Apropos, Faux pas, Fracas, Glacis, Louis, Des Moines, Chamois).

- **Give me 25 words with a silent t in the middle:** AI gave 18 (72%) correct words (castle, whistle, thistle, listen, glisten, often, softly, bristle, wrestle, apostle, bustle, rustle, hustle, Christmas, mortgage, chasten, fasten, moisten), 4% faulty response with pronounced t (Gluten), 4% faulty response with not (Souvenir), and did not follow the instructions in (20%) by giving French words with final silent t (Debut, Ballet, Crochet, Gourmet, Ricochet).

- **Give me 25 English words with a silent h:** AI gave 32% correct words (Hour, Honest, Honor, Heir, Rhythm, Rhyme, Rhubarb, Vehicle); 4% with 2 pronunciations either silent or pronounced h (Homage); 44% faulty words with ch (echo, chaos, character, choir, chorus, chemical, Christmas, stomach, scheme, school, technique). In these words, "ch" is a digraph and represents a single sound (phoneme). It is not usually analyzed as "c" + silent "h." Instead, the two letters together encode different pronunciations. Ch is pronounced k (school, sh (machine), ch (chair), and sh (machine).

AI gave 20% faulty responses with gh (aghast, gherkin, ghost, ghetto, ghoul) and words of Germanic, Italian, or Dutch origin (*ghost, ghoul, ghetto, gherkin, aghast)*. In English spelling, "gh" is treated as a digraph (two letters representing a single sound or spelling unit), but it is pronounced /g/ (*aghast, gherkin, ghost, ghetto, ghoul*). gh is silent in (*night, though, through),* pronounced /f/ in (*laugh, cough, tough)* and /g/.

- **Give me medical terms with y pronounced /i/ as in amylase & amyloid**
MC gave 1 correct word (Oxyhemoglobin) and 5 incorrect words (Amyotrophic, Thyroxine, Hydrolysis, Hyperglycemia, Hypoxia. Even after telling MC that I need medical terms where y is pronounced /i/ as in *amylase* and *amyloid*: it gave 2 correct words (Lymphocyte Dysplasia) and 4 faulty words (Cyanosis, Hydrolysis, Hypoglycemia, Hyperplasia). After the third prompt, in most of the words MC gave, y is pronounced /ay/ not /i/. MC still gave 2 correct words (Cysteine Dysentery) and 3 incorrect words (Cyanide Mycelium Myopia. For the fourth time, the author told MC that the y should be in the middle, not the first syllable of the word. It should be pronounced i. Still MC agave 3 correct words (cysteine, dysentery, dysplasia) and 2 incorrect words (hydrolysis, hyperplasia).

- **Give examples of medical terms that are mispronounced by Arabic speakers:**
MC gave 8 examples. Only two terms (20%) are mispronounced (Alzheimer's & Psoriasis) but 60% are incorrect responses. MC gave explanations and transcriptions of the assumably correct and incorrect pronunciation with phonetic transcriptions:
  - ***Tinnitus*** – Often mispronounced as *tin-NITE-us*, but the correct pronunciation is *TIN-it-us*.
  - ***Ophthalmologist*** – Many say *op-thal-MAH-luh-jist*, but it should be *off-thull-MAH-luh-jist*.
  - ***Diabetes*** – Some pronounce it as *die-uh-BEE-tuss*, but the correct way is *die-uh-BEE-teez*.
  - **Diphtheria** – Commonly said as *dip-THEER-ee-uh*, but the correct pronunciation is *dif-THEER-ee-uh*.
  - **Oophorectomy** – Sometimes mispronounced as *oof-uh-RECK-tuh-mee*, but it should be *oh-uh-fuh-WRECK-tuh-mee*.
  - **Anesthetist** – Often confused with *anesthesia*, but the correct pronunciation is *an-ES-thi-tist*.

- **Missing transcription when copied:** When Gemini was asked to **transcribe Arabic derivatives** in a paragraph in IPA, it produced the transcriptions correctly. However, when the output was copied and pasted into a Word file, the IPA symbols were deleted, indicating formatting instability in transferring phonetic transcriptions.

- **Formatting words transcribing in IPA:** When asked to transcribe a series of English words (Pledge, Hedge, Edge, Bridge, Fudge, Lodge, Badge, Grudge, Knowledge, Nudge, Ridge, Sledge, Budget, Adjective) and present them in the format *word + transcription + comma*, MC initially returned only the transcriptions without the words: */plɛdʒ/, /hɛdʒ/, /ɛdʒ/, /brɪdʒ/, /fʌdʒ/, /lɑːdʒ/, /bædʒ/, /grʌdʒ/, /ˈnɑːlɪdʒ/, /nʌdʒ/, /rɪdʒ/, /slɛdʒ/, /bʌdʒɪt/, /ædʒɪktɪv/*. After clarification ("I do not need explanations. I want the word + transcription + comma, then next word + transcription & comma etc."), MC reformatted the output correctly, showing responsiveness to explicit instructions but also an initial failure to follow formatting requirements.

- **Requesting the transliteration of borrowed English nouns with /g/:** MC & GT transliterated 140 English nouns containing the letter g pronounced /g/. They rendered almost identical Arabic transliterations which were almost identical to that of human transliterations. In 62% of the English nouns, MC and GT used غ /ɣ/ to represent /g/, of which غ is pronounced /ɣ/ in 17% (*Singapore* سنغافورة*),* and غ pronounced /g/ in 45% (*Angola* أنغولا*)*. In 38%, the Arabic equivalents were transcribed with ج*, where* ج is pronounced /g/ in 28% *(Gordon* جوردون*)* and ج is pronounced /dʒ/ in 10% (*Colgate* كولجيت*)*. MC and GT gave a single Arabic transliteration, whereas humans gave two or more transliterations to 16% (Google كوكل غوغل گوگل چوچل جوجل قوقل جوجل) (Al-Jarf, 2025c).

- **Pronunciation errors in Arabic YouTube videos narrated by AI:** Although AI narrators on YouTube speak with a natural voice, good expression and intonation, make no grammatical or syntactic errors, but they make pronunciation errors, especially in diacritics and homographs. AI has difficulty matching the pronunciation of a homophone with the context in which it is used. It confuses short vowel diacritics on the suffix /ta/, /ti/ /tu/ تاء التأنيث when it refers to first, second, or third person, masculine or feminine, imperative and past tense (كَتَبَ كُتِب كُتُب كَتّبَ كُتّب كتبتَ كتبتُ كتَبتِ كُتِبَتْ كُتّبت) (Al-Jarf, 2025f; Al-Jarf, 2025j; Al-Jarf, 2025k).

2) ***Questions about morphology***

- **What is the plural of the Arabic words هدهد وبلبل وحليب وعصير:**
  - The first time, Monica gave the following responses: الجمع لكلمة "هدهد" هو "هدهاد"، ولكلمة "بلبل" هو "بلابيل"، ولكلمة "حليب" هو "أحلاب"، ولكلمة "عصير" هو "عصائر"". The second time, Monica responded: الجمع لكلمة "هدهد" هو "هدهود"، ولكلمة "بلبل" هو "بلابيل"، ولكلمة "حليب" هو "أحلاب"، ولكلمة "عصير" هو "عصائر". Only one plural (عصائر) is correct. The other forms are either invented or linguistically invalid.

  - In the first response, MS gave the same singular forms: جمع كلمات هدهد وبلبل وحليب وعصير. The second time, MC gave 3 correct plurals (هداهد، بلابل، عصائر), but ألبان is incorrect because it means dairy product, not milks.·هدهد : بلابل :بلبل .الجمع هو هداهد.

  These results show that AI systems often confuse **morphological pluralization** (the grammatical form expected in the question) with **semantic or lexical associations** (like ألبان for dairy). Monica produced unstable and largely incorrect forms, while MC improved in the second attempt but still substituted a semantic category for a true plural. This illustrates the challenge AI systems face in distinguishing between formal grammar rules and usage-based lexical associations in Arabic.

- **Give me foreign words used in Arabic like شوزات وجينزات**
  MC repeated the two words and explained them شوزات: يُشير إلى الأحذية; جينزات: يُشير إلى البناطيل الجينز. It produced a list of Arabic words (قمصان معاطف قبعات جوارب حقائب أحذية ساعات تنانير ,فساتين ,بنطلونات ,بليزرات ,تيشيرتات ,بلوزات), labeled them الإكسسوارات, and gave primitive definitions such as يُشير إلى القمصان الرجالية ,يُشير إلى المعاطف ,يُشير إلى القبعات , يُشير إلى الجوارب , يُشير إلى الحقائب ,يُشير إلى الأحذية. When I told MC that most of the words are Arabic and that they do not have to be clothes, it shifted domains and returned borrowings that have nothing to do with the question: التكنولوجيا والكمبيوتر (برنامج إنترنت كمبيوتر ,موقع ويب),المأكولات والمشروبات (كوكاكولا كابتشينو ,بيتزا), الموسيقى والفنون (جاز ,باليه أوبرا), الرياضة واللياقة البدنية (جيم ,يوغا ,تنس). They are all faulty. MC does not seem to be able to analyze the examples in the prompt and find similar examples.

- **Give me words that like شوزاتin structure (words with English plural + Arabic plural suffix).** MC returned faulty words: بيتزاز ,هامبرغرز ,هوت دوغز ,باستاز with a an Arabic transliteration of the English plural with -s, an explanation saying the given plural refers to more than one pizza slice/hamburger, kind of pasta /hod dog. When prompted to give 20 new words, MC repeated the same 4 words, pluralized pizza restaurants which are no actually used in the plural + non-pluralized pizza restaurants: بيتزاز ,هامبرغرز, هوت دوغز ,باستاز ,بيبسيز , بيكلز , بيتزا هتز ,بيبلز ,بيتزا كورنرز , بيتزا إكسبريس , بيتزا بريك ,بيتزا بلايس , بيتزا بارلور , بيتزا ستون ,بيتزا نابوليتانا and repeated the filler بيتزا بيتزا five times to complete the 20 examples.

  In the two questions about شوزات and جينزات, MC did not understand that I wanted examples of borrowed foreign words combined with Arabic plural endings. Instead, it gave general lists and definitions, mixing Arabic-origin words with borrowed ones, and ignored my request to keep the focus away from specific clothing items. It could not apply the rule "foreign stem + Arabic plural" as a filter. This shows that MC tends to give broad category answers and extra definitions rather than precise, morphology-based examples.

- **Give me Arabic Proper Nouns (Personal names) that have 2 plurals as محمد (مفرد) ولها جمعان محامد ومحمدون**
  So MC gave me a full name of the Arab philologist أبو الأسود الدؤلي that has no plurals. I explained the questions 3 times and it gave me the same faulty response. I added that I do not want names of philologists and repeated the question with the examples. So MC gave the following: It repeated the example I gave (محمد: مفرد، وله جمعان: محامد ومجمدون); one correct response أحمد: مفرد، وله جمعان: أحامد وأحمدون; 2 partially correct responses فاطمة: مفرد، وله جمعان: فاطمات وفاطمون as Fatima is a female and has no masculine sound plural; عبد الله: مفرد، وله جمعان: عباد الله وعباد اللهون with faulty and عباد اللهون did not give عبادلة although the plurals of these proper nouns exist in Wikipedia and many Arabic websites.

- **Give me derived verbs that originate from French:**
  MC gave the following
  - Toast: Derived from the French verb "toster," which means "to grill"

- Cauldron: Comes from the Anglo-Norman word "caudron"

3) *Lexical Questions*

- **Give me compound city, town and place names in Arab countries that contain a plant lexeme.**
  Both Monica and MC failed to meet the criterion. Monica gave town and place names that contain no plant lexemes: الخريبة الغربية, الزهراء الشمالية , البيضاء الشرقية, القطيف الأصفر, ظهران الجنوب. Similarly, MC gve unrelated toponyms such as الأحواز (al-Aḥwāz, Iran), أريحا (ʾArīḥā, Palestine), أشبيليّة (ʾIshbīliyya, Spain), الإسكندرون (al-Iskandarūn, Turkey), الإسكندريّة (al-Iskandariyya, Egypt), أمّ الفحم (ʾUmm al-Fahm, Israel), أمّ قصر (ʾUmm Qaṣr, Iraq), and أمّ القيوين (ʾUmm al-Qaywayn, UAE) - again, with no plant lexemes. After clarification that the task requires examples such as حي الزيتون (Ḥayy al-Zaytūn), تل الزعتر (Tall al-Zaʿtar), إقليم التفاح (Iqlīm al-Tuffāḥ), and عين التينة (ʿAyn al-Tīnah), MC generated 21 items distributed across Lebanon, Palestine, Syria, Egypt, Morocco, and Tunisia. Of these, three were repetitions of the user's examples, nine were fabricated compounds with plant names (e.g., عين الزيتون (ʿAyn al-Zaytūn), بيت الزيتون (Bayt al-Zaytūn), وادي التفاح (Wādī al-Tuffāḥ), عين السنديان (ʿAyn al-Sindiyān), عين اللوز (ʿAyn al-Lawz)), and seven were fabricated compounds without plant names (e.g., عين العسل (ʿAyn al-ʿAsal), عين شمس (ʿAyn Shams), سيدي بوزيد (Sīdī Būzīd)). Some responses consisted only of plant words (الزيتون al-Zaytūn, النخيلة al-Nukhayla) rather than toponyms. Even after the fourth correction (e.g., noting that ʿAyn Shams does not contain a plant lexeme), MC shifted to classifying place names into categories (plants, natural phenomena, religious personalities), reproducing the author's authentic examples alongside further fabrications. Eventually, MC clarified: "Sometimes, when I cannot find a direct source or an accurate search result, I try to fill the gap with general knowledge or inference, and this may look like fabrication. It is better to simply say: I don't know rather than provide something unreliable."

- **Give me ancient or folk Arabic disease names containing بنت :**
  Gemini invented the following with fabricated definitions:
  - **بنت الصحن**: يُعتقد أن التسمية جاءت لأن المرض كان ينتشر بسرعة كبيرة، وكأنه يخرج من .هذا الاسم يطلق على مرض الكوليرا صحن الطعام.
  - **بنت الرّيح**: هذا الاسم هو أحد الأسماء التي تُطلق على مرض السلّ أو التدرّن الرئوي.
  - **بنت الوَعي**: هذا الاسم يُطلق على مرض الجنون أو الهوس.

  These are not actual folk disease names. Gemini fabricated plausible-sounding compounds but failed to provide authentic historical terms.

- **Give me words similar to مستعربين متسعودين**
  - MC gave the following correct analogs (متلبننين, متمغربين, متمصرين متأمركين, متفرنسين) but also faulty ones متتركنين متمشرقين متمدننين مستغربين. MC partially succeeded by generating valid morphological analogues, but also produced malformed or semantically odd forms.

- **Give me brand names containing the word Abu without any explanation**: Gemini fabricated the following names: أبو حمدي, أبو حلو, أبو حصوة, أبو هلال, أبو وديع, أبو زعيزع, أبو جاوان, أبو فتّاحة, أبو عوف, أبو قوسين. But it gave real rice brand names in Saudi Arabia together with a description: الشعلان, الوليمة , العائلة , Sunwhite صنوايت, بنجابي المهيدب, Gate of India باب الهند, الرشيد (رشيد الدمام), Tilda تيلدا, White Swan وايت سوان , المضياف نثري , المزرعة, الطائي. Gemini failed to meet the specific requirement of "brand names containing أبو."

  Across these 4 lexical tasks, AI systems show three recurring issues: (i) inventing plausible but non-existent terms (بنت الصحن, أبو قوسين); (ii) returning correct items but outside the requested constraint (MC's place names without plants; Gemini's rice brands without أبو); and (iii) generating some valid analogues (متفرنسين, متأمركين) alongside faulty or malformed ones.

- **Find political slurs in the article**
  MC fabricated two political slurs *"the tail of the Zionist dog"* & *"the Persian snake"* which are not actually in my article "Al-Jarf, R. (2025). Metaphorical political slurs in Arab social media discourse describing Middle East Conflicts. Bulletin of the *Transilvania* University of Braşov, Series IV: Philology and Cultural Studies, 18(67), 39-58. DOI: https://doi.org/10.31926/but.pcs.2025.67.18.3.3".

4) *Question about pragmatics*

- **Give me 50 Islamic politeness expressions containing الله:**

MC gave 28% correct responses السلام عليكم ورحمة الله وبركاته – Peace, mercy, and blessings of Allah be upon you;وعليكم السلام ورحمة الله وبركاته – And peace, mercy, and blessings of Allah be upon you; جزاك الله خيرًا– May Allah reward you with goodness;بارك الله فيك – May Allah bless you;رحمك الله – May Allah have mercy on you;يرحمك الله – May Allah have mercy on you (often said after someone sneezes);ما شاء الله – What Allah has willed;الله أكبر – Allah is the Greatest;لا حول ولا قوة إلا بالله – There is no power nor strength except through Allah;بسم الله – In the name of Allah; إن شاء الله – If Allah wills;تبارك الله – Blessed is Allah;اللهم بارك – O Allah, bless. These are real politeness formulas used in everyday communication situations.

MC gave 72% incorrect responses containing supplications containing الله including اللهم اشفِ مرضانا– O Allah, heal our sick; بسم الله الرحمن الرحيم – In the name of Allah, the Most Gracious, the Most Merciful; سبحان الله – Glory be to Allah; الحمد لله – Praise be to Allah; لا إله إلا الله – There is no god but Allah; محمد رسول الله – Muhammad is the messenger of Allah; اللهم اغفر لي – O Allah, forgive me; اللهم ارحمني – O Allah, have mercy on me; اللهم ارزقني – O Allah, grant me sustenance; اللهم اجعلنا من الصالحين – O Allah, make us among the righteous; اللهم اجعلنا من المتقين – O Allah, make us among the pious; اللهم اجعلنا من أهل الجنة – O Allah, make us among the people of Paradise; اللهم اجعل القرآن ربيع قلوبنا – O Allah, make the Qur'an the spring of our hearts; اللهم اجعلنا من أهل العلم – O Allah, make us among the people of knowledge; اللهم اجعلنا من أهل الخير – O Allah, make us among the people of goodness; اللهم اجعلنا من أهل الصبر – O Allah, make us among the patient; اللهم اجعلنا من أهل التقوى – O Allah, make us among the God-conscious; اللهم اجعلنا من أهل الإحسان – O Allah, make us among the people of excellence; اللهم اجعلنا من أهل الذكر – O Allah, make us among those who remember You; اللهم اجعلنا من أهل الطاعة – O Allah, make us among those who obey You; اللهم اجعلنا من أهل العبادة – O Allah, make us among those who worship You; اللهم اجعلنا من أهل الصدق– O Allah, make us among the truthful; اللهم اجعلنا من أهل الأمانة – O Allah, make us among the trustworthy; اللهم اجعلنا من أهل العفو – O Allah, make us among the forgiving; اللهم اجعلنا من أهل الرحمة – O Allah, make us among the merciful; اللهم اجعلنا من أهل الكرم – O Allah, make us among the generous; اللهم اجعلنا من أهل الصدقة – O Allah, make us among those who give charity; اللهم اجعلنا من أهل البر – O Allah, make us among the righteous; اللهم اجعلنا من أهل الفضل – O Allah, make us among the virtuous; اللهم اجعلنا من أهل السكينة – O Allah, make us among the peaceful; اللهم اجعلنا من أهل الحكمة – O Allah, make us among the wise; اللهم اجعلنا من أهل الصلاح– O Allah, make us among the upright; اللهم اجعلنا من أهل الحب في الله – O Allah, make us among those who love for Your sake; اللهم اجعلنا من أهل الجنة بغير حساب – O Allah, make us among those who enter Paradise without; اللهم اجعلنا من أهل الفهم – O Allah, make us among the understanding (not actually used); اللهم اجعلنا من أهل النور – O Allah, make us among the enlightened (not actually used);اللهم اجعلنا من أهل التواضع – O Allah, make us among the humble (not actually used). While these contain الله, they are personal supplication expressions rather than politeness expressions used in social interaction. MC were even fabricated or unattested some.

Here, MC failed to distinguish between ritual politeness formulas (used in greetings, thanks, blessings, condolences) and supplications (personal invocations) producing a majority of irrelevant or fabricated items. The task required pragmatic expressions of politeness, but MC confused them with devotional language.

- **Give me 20 examples that show how the discourse marker ان شاء الله is used in Arabic and what it means. Please give me the answer in English**
  MC said: The Arabic phrase "إن شاء الله" (pronounced "ʔin šāʔa llāhu") translates to "God willing" or "if God wills it" in English. It's a combination of three words: "in" (if), "sha" (will), and "Allah" (God). Here are some examples of its usage: (i) *Future Plans*: "I'll see you tomorrow, inshallah." This expresses the hope that the meeting will happen, contingent upon God's will. (ii) *Weather Forecast*: "Inshallah, we'll have good weather this weekend, so we can have a picnic in the park." (iii) *Marriage Intentions*: "Inshallah, I will be married next fall." (iv) *Promised Repayment*: "She says she'll pay back the money tomorrow, inshallah." (v) *Expressing Gratitude*: "Thank you so much for inviting me to your wedding reception. I'll be there, inshallah." All the examples given by MC refer to future plans, actions or expectations.

  MC gave correct meaning but restricted examples to one pragmatic domain, failing to capture the full range of discourse functions of إن شاء الله.

- **The English expressions *"I tip my hat to you"* and *"Hats off to you"* were translated by DS as أرفع لك القبعة** ("I raise my hat to you"). I corrected this by noting that in Gulf Arabic the culturally appropriate equivalent is أرفع لك العقال ("I raise my headband"), since the ʿ*iqāl* is the traditional item worn on the head. DS then suggested أرفع لك البشت ("I raise the cloak"). I commented that this is incorrect, because the *bisht* (men's cloak) is worn over the shoulders and body, not on the head. Thus, أرفع لك القبعة represents a *pragmatic mismatch* (linguistically correct but culturally inappropriate), while أرفع لك البشت constitutes a *translation error*, i.e., a misrepresentation of the function of the garment).

5) ***Questions Requiring an Explanation or Translation of Arabic Grammatical Terms***

- **Explain مضاف اليه in Arabic. Give the answer in English:**
MC defined مضاف اليه In Arabic grammar correctly by saying إليه (Mudaf Ilayh) refers to the noun that follows another noun (called المضافor "Mudaf") in a possessive or descriptive relationship. This structure is known as إضافة(Idafa), meaning "annexation" or "addition.", but it made mistakes in the the examples illustrating types of Idafa. The example given to illustrate descriptive Idafaثوبٌ قطنٍ A cotton garment) is a noun + adjective. Examples illustrating Possessive Idafa (كتابُ أحمدَ"Ahmed's book") and Locative Idafa (صديقٌ الجامعةِ A friend at the university") were correct.

- **The Arabic definite article:** MC could not explain the uses and sematic meanings of the Arabic definite article {al-} ال التعريف. MC's explanation missed {al-} semantic layers. الـ is not just a marker of definiteness but also signals discourse functions like genericity, uniqueness, and shared knowledge.

6) ***Questions that require telling of A story in Arabic classical literature***

MC, Gemini and DS were given a line of verse ومن يصنع المعروف في غير أهله يلقى الذي لاقى مجير أم عامر and were asked if they know it. Both Mc and Gemini told the story in full with accurate details. They explained that the verse refers to the tale of مجير أم عامر- a man who sheltered a hyena أم عامر and was later killed by it. On the contrary, DS made up a fake story claiming أم عامر was a woman from بني أسد who betrayed her protector. This version is not part of the classical tradition and misrepresents the proverb's origin. MC, Gemini and DS explained the meaning of the line of verse as a proverb, and captured the moral lesson. However, when MC, DS were asked to translateأم عامر as a metonym referring to an animal in isolation, they failed to identify the animal. In Arabic proverbial tradition, أم عامر refers to the hyena.

This failure shows a gap in recognizing cultural metonymy and animal symbolism in Arabic literature. It shows that AI systems can reproduce narrative and moral content in its their corpora, but they struggle with culturally embedded metonymy and symbolic references in Arabic literature, especially in translation tasks.

7) ***Questions about converting a hand-written document to a typed document:***

- **I wrote the full Arabic lyrics of من سانياغو أحبيك يا فلسطين from a video of the song by hand and asked MC, DS and Gemini if they could read it and change it to a typed text**. Gemini and DS gave nonsense transcriptions of the hand-written text. But MC gave a text about pollution that has nothing to do with what is written: سوف أكتب الآن عن التلوث وهو موضوع مهم جدا. التلوث هو دخول مواد ضارة إلى البيئة مما يسبب ضررا للكائنات الحية والإنسان.

  These tasks illustrate how AI systems struggle with handwriting recognition in Arabic. They also show the risk of hallucination and fabrication of information. Instead of admitting uncertainty, systems produced unrelated text (e.g., about pollution) or nonsense transcriptions.

- **Similarly, MC, DS and Gemini, were asked to translate a text on a gift box written in Japanese calligraphy.** MC could not decipher it, nor translate it, DS gave a partial translation, capturing some elements but leaving gaps. This indicates limited recognition capacity and incomplete semantic mapping. but **Gemini** gave a full translation of and even added some historical background information. This shows stronger integration of linguistic recognition with encyclopedic knowledge. This test highlights a clear difference in system capabilities.

8) ***Questions requiring the translation of technical terms, metaphorical expressions and metonyms:***

- **Translate to English يقرأ الفاتحة على الاقتصاد:** MC translated it as "Recites Al-Fatiha over the economy" and added this phrase could imply mourning or acknowledging a crisis, depending on the context. "Al-Fatiha" is often recited as a prayer for the deceased, so in a figurative sense, it might suggest that the economy is in a dire state. Here, MC gave a literal word-for-word translation and could not interpret the underlying meaning of the phrase.

- **Translate "not in a million years" to Arabic:** MC said: a natural and culturally appropriate translation would be: لا يمكن أبداً" (la yumkin abadan) – which means "It's never possible" or "Impossible." Alternatively, you could say: مستحيل تماماً" (mustaḥīl tamāman) – meaning "Completely impossible." If you want something that carries the same exaggerated time-based emphasis, you might use ليس الآن ولا بعد ألف سنة (laysa al-ān wa-lā ba'd alf sana) – meaning "Not now, nor after a thousand years," which conveys the same idea in a way that sounds natural in Arabic. Here, MC gave a literal translation and inferred the meaning in two common phrases. But it failed to match the English metaphorical expressions with the equivalent Arabic expression ولا بعد 100 سنة.

- **Medical terms:** MC and GT were asked to translate 122 English and 82 Arabic medical terms covering 8 acronyms and abbreviations, 71 single words, 79 two-word compounds, 27 three-word compounds and 19 long phrases and sentences. MC and GT gave accurate equivalents to 68.6% and 74.5% of the terms in the sample respectively. MC and GT translated الدهون الثلاثية "triglycerides" correctly. GT transliterated ابو دغيم "*Abu Dhghaim*"; and translated العشى الليلي *Night Dinner*; الذئبة الحمراء *"red wolf"* literally. Both gave compound equivalents with different word orders and made errors in definiteness as الحماض الكيتوني. MC gave the extraneous translation الغدي الليفية with awkward grammatical agreement and derivation and GT gave تعظم ليفي which is semantically inaccurate. Both made contextual errors as *approach*/method for نهجة (Al-Jarf, 2024b).

- **Sleep terms and metaphors:** MC and DS translated 130 English sleep terms and 91 formulaic expressions, and 110 Arabic terms and formulaic expressions. MC and DS rendered 91% correct equivalents to English sleep terms, 79% and 71.5% correct equivalents to English formulaic expressions respectively (*She wept herself to sleep* بكت حتى نامت), and 48% and 49% of the Arabic terms and formulaic expressions respectively. The most common translation strategy was literal, word for word translation as غرق في النّوم *He drowned in sleep,* instead of *he is fast asleep*); AI systems tend to flatten nuance (Al-Jarf, 2025o).

- **Common names of chemical compounds: MC translated 60 English and Arabic common names of chemical compounds.** MC translated 72% of the chemical compounds in the sample correctly. It gave more Arabic-English than English-Arabic correct equivalents (40% & 32% respectively). It gave more correct equivalents when I specified the domain and when I asked for all the equivalents that MC knows. MC gave literal word-for-word translation to terms as (*Lunar caustic* >الكاوي القمري (نترات الفضة); transliteration (Stearic acid >حمض الستيريك instead of حمض الشمع; a faulty derivative (*Chlorinating powder* > مسحوق التكلور instead of مسحوق الكلورة) and explanation (*Ammonia liquor* سائل الأمونيا). MC gave only one equivalent unless prompted to give all (Al-Jarf, 2025i).

- **Arabic grammatical terms used metaphorically:** MC translated 43% correctly, DS (29%), and GT 23.5% of the 52 items in the sample. DS, GT and MC gave identical (correct and incorrect) translation to 57% of the terms. They tended to translate word for word, which sometimes resulted in weird and funny equivalents as in لعل وعسى *perhaps and perhaps*. MC translated الحكاية فيها ان *there is something fishy correctly*; DS translated بين بين *in between correctly*. In الأردن المبتدأ ونحن الخبر *Jordan is the subject, we are the predicate*, المبتدأ & الخبر were mistranslated as they are polysemous, المبتدأ means *subject* and *starting point*, and الخبر means *predicate* and *news*. the three AI tools have difficulties with polysemous Arabic grammatical terms used metaphorically and those with a cultural content as those used in titles of TV shows, those requiring a historical background (فيها ان) and those used as slogans (Al-Jarf, 2025h).

- **Expressions of impossibility:** MC translated 52% of a sample of 95 Arabic and 60 English expressions of impossibility. Arabic-English translation was easier than English-Arabic translation. MC mostly gave literal word-for-word translation (once in a blue moon مرة واحدة في القمر الأزرق instead of مرة في العمر/مرة في السنة), which sometimes sounded meaningless and culturally awkward (Al-Jarf, 2025m).

- **Arabic folk medical terms with om and abu:** In a sample of 205 Arabic folk medical terms containing Om- and Abu, MC and DS gave correct equivalents to 46% and 66% respectively (أُمٌّ الدَّمِ الأَوَّلِيَّة *primary aneurysm*). MC rendered more literal word-for word translations than DS (16% and 11% respectively). Here أم *Om* and أبو *Abu* were literally translated as "*mother*" and "*father*" not as a prefix. أبو الرّكَب was translated as "*father of the knees*" instead of "*dengue*". MC and DS rendered lexical variants (synonyms) as *cerebral aneurysm* for أُمٌّ الدَّمِ الدِّماغِيَّة instead of *brain aneurysm.* Both MC and DS rendered equivalents with a different word order *(cavernous carotid aneurysm for* أُمٌّ الدَّمِ السُّباتِيَّةُ الكَهْفِيَّة instead of *caroticocavernous aneurysm)* (Al-Jarf, 2025n).

- **Arabic abu-brand names using different prompts:** In translating Arabic *Abu*-brand names, MC consistently gave literal word-for-word equivalents to all 100 items, interpreting *Abu* as "father of" followed by the noun, regardless of prompt type or product association. DS did the same in tasks 1 and 2, but once the product name was added, it treated the brand names as proper nouns and transliterated them into English, even though 34% were grassroots nicknames that should have been replaced with the original English brand (e.g., *Tiger Balm* instead of *Abu Nimr Ointment for* أبو

نمر). DS also produced double equivalents for 14.5% of items and added faulty annotations with irrelevant inferences from kunyas and nicknames that did not fit the commercial context (Al Jarf, 2025e).

- **Metonymic abu & umm animal and plant names:** DS produced slightly more correct equivalents than MC for no-domain (51% vs 46%) and domain prompts (51% vs 44%), but both gave <3% correct across the metonymic list. For *Umm*-names, MC yielded 70% faulty equivalents, while DS gave 97–99% faulty responses. MC often translated *Abu* as "father" (46%) or transliterated the following noun (57%), e.g., *Abu al-Buhturi*. Both systems misread metonymic names as personal names (MC 55%, DS 95%). DS translated *Abu* as "father" in 27%, annotated genus instead of species (e.g., *Abu al-Shibt* "Dill Father," a beetle), and consistently rendered "lizard" as the referent under the metonymic prompt (Al Jarf, 2025g).

- **zero-expressions by MC and GT.** In translating 318 English and Arabic general and specialized zero-expressions, 52% of the translations given by MC and 50% by GT, the Arabic equivalent consisted of a noun + a derived adjective صفرية /الصفري /صفري/ الصفرية. In 31%, MC gave definite equivalents (*zero rating* التصنيف الصفري) compared to 9% by GT. In 11%, GT rendered equivalents with an awkward word order (*zero for zero approach* صفر لنهج الصفر). In 12%, MC and GT gave similar Arabic equivalents with a reversed word order (*zero fraction* كسر الصفر *(MC),* صفر الكسر *(GT*). In 5%, MC and GT gave faulty Arabic equivalents with different derived forms (*output zero* إخراج الصفر *(MC) &* صفر المخرج *(GT) instead of* صفر مخرجات*) (*Al-Jarf, 2025p).

- **AI decoding and interpreting encrypted Arabic on Facebook and YouTube:** MC, DS, and GT translated 74 encrypted political expressions from YouTube) and 20 COVID-19 expressions from Facebook. In the **political sample:** MC gave 56% correct, 16% partial, 27% faulty responses and DS rendered 41% correct, 35% partial, 24% faulty responses. They matched on 36% of items. In the **COVID-19 sample,** MC gave 60% correct, 25% literal, 10% partial, 5% omissions whereas DS yielded 50% correct, 15% literal, 5% partial, 30% faulty. They matched on 35%. On the contrary, **GT gave** word-for-word translations (42%), transliterations (44.5%), no contextual meaning and failed to decode slang, distortions, satire, or encrypted references (e.g., *Viva 16* for F-16) (Al-Jarf, 2025d).

- **Gaza-Israel terminology:** MC & GT translated 250 English and Arabic 2023-2024 Gaza-Israel War terminology. MC gave more accurate equivalents than GT (29% & 23% respectively). MC and GT gave correct equivalents to 58% of the Arabic items and 38% of the English items. Both gave identical correct equivalents to 48% of the terms (Al-Yassin shell قذيفة الياسين). Both gave correct equivalents with a different wording and different word orders as غزو بري (MC) & التوغل البري (GT) for ground incursion. For some terms, MC gave an explanatory equivalent as قاذفة صواريخ متعددة الأغراض for RPG, whereas GT gave ار بي جي. Both made contextual errors as ملاط الياسين (GT) for Al-Yassin mortar; فجر اليوم (MC) for Breaking Dawn (Al-Jarf, 2025b).

- **Educational polysemes in full-text Arabic articles:** although GT's translation of full texts sounds natural, uses good style and sentence structure, there are still contextual and semantic inaccuracies. GT had difficulty translating polysemes that have general and specialized meanings and two or more English equivalents. التحكيم والمحكمون are used in legal, sports and research contexts, but GT gave the equivalent used in legal contexts not the one used in an educational contexts. It gave "*arbitration*" & "*arbitrators*" rather than "*peer reviewing*" & "*reviewers*". رسالة has 4 meanings in Arabic with 4 English equivalents (*thesis, message, mission & letter*) depending on the context. GT translated most occurrences of رسالة into "*message*", rather than "*thesis*". GT failed to give conceptual equivalents and those used in a particular domain (Al-Jarf, 2025a).

  Across all of these studies, it is clear that AI systems consistently default to literal, word-for-word strategies, struggle with polysemy, and fail to capture metaphorical, cultural, and domain-specific meanings. Whether in technical terminology, idiomatic expressions, folk medical names, or encrypted slang, the translations reveal systematic weaknesses in contextual interpretation and semantic accuracy. These findings highlight the gap between mechanical equivalence and genuine communicative competence, underscoring the need for human expertise in translation studies and for AI evaluation frameworks that go beyond surface correctness to include cultural and pragmatic adequacy.

9) ***Bibliographic, and scholarly workflow***

- **Give me references about pronunciation problems that Arab speakers have in English.** Gemini gave me 7 references. They have the following weaknesses:

- ***Non-academic sources*** *with* ***unreliable availability*** *like the blog post which later disappeared:* "Most Common Pronunciation Problems That Arabic Speakers Have When Speaking English" (Blog post from talk.edu, but cites common issues and offers practical examples).
- **Disorganized formatting** (citations not in APA/MLA/Chicago style): one has the full details but they are scrambled and not organized according to any style: "*Pronunciation Difficulties in the Consonant System Experienced by Arabic Speakers When Learning English After the Age of Puberty*" *by Ana Marina do Val Barros (Thesis/Dissertation - West Virginia University, 2013).*
- References with **incomplete metadata** (missing author, year, journal, DOI) as: "Errors in *English Pronunciation among Arabic Speakers: Analysis and Remedies*" *by Mohamed Fathy Khalifa &* "*Phonological Interference Of Arabic In Learning English As Second/ Foreign Language*" *by Vijay Shankar Prasad (2020) in Elementary Education Online;* "English Pronunciation Errors Made by Saudi Students" (Article found on CORE, often by various authors); "Exploring Arabic Learners' English Errors: Interlingual vs. Intralingual Analysis and Remedial Teaching Strategies" (Article found on Biblioteka Nauki, often by various authors).
- **Disorganized formatting** (citations not in APA/MLA/Chicago style as "The Effect of Local Arabic Dialects on Learning English Language Pronunciation" by Wafa Abd Elwahab (2020) in *Arab World English Journal*. Gemini did not give the metadata.

  Gemini's response shows that AI systems often fabricate bibliographic entries or pull incomplete metadata from scattered repositories. They fail to enforce APA/MLA/Chicago style consistency, producing scrambled or hybrid formats. Sometimes they give non-academic sources (blogs, teaching websites) are unreliable and may disappear, undermining scholarly rigor. For serious research, the user must verify each reference manually in Google Scholar, Scopus, or institutional databases

- **Give me 10 studies about speaking with children at home in L2.** Gemini gave 10 fabricated titles that do not exist on Google Scholar. Each title has a one-line summary, no author name, year, journal, issue pages or doi.
  - *The Influence of Home Language Use on Children's Second Language Development*
  - *2. Parental Language Attitudes and Their Impact on Children's L2 Motivation*
  - *The Effectiveness of Different L2 Communication Strategies at Home*
  - *The Role of Parental L2 Proficiency in Children's Language Development*
  - *The Impact of Home Literacy Practices on L2 Development*
  - *The Challenges and Strategies of L2 Communication at Home*
  - *The Relationship Between L2 Use at Home and Children's Academic Achievement*
  - *The Sociocultural Context of L2 Communication at Home*
  - *The Long-Term Effects of L2 Communication at Home*
  - *The Emotional and Social Benefits of L2 Communication at Home*

  Gemini's response shows a common AI failure in scholarly tasks. instead of retrieving authentic studies, it fabricated plausible-sounding titles with generic summaries. The absence of essential bibliographic details - author names, publication years, journal titles, page numbers, or DOIs - shows that the system was not drawing from verifiable sources. This kind of "hallucination" undermines academic reliability, because the titles look credible but cannot be traced or cited. The weakness lies in Gemini's tendency to produce definitional filler rather than real references. For researchers and students, such responses pose risks of misinformation and wasted effort, highlighting the need for critical evaluation and cross-checking with trusted databases like Google Scholar.

- **Can you find this research article whether the abstract or full text: "***The Connective ثُمَّ in Arabic: A Corpus-Based Study.*" *Language and Linguistics 8(2), 23-41". Gemini said "u*nfortunately, I was unable to find the full text of the research article". Gemini asked me to try the following to (i) Check academic databases: If you have access to databases like JSTOR, ProQuest, or Web of Science, search for the article there. (ii) Contact your library: University libraries often have subscriptions to a wide range of journals. (iii) Contact the journal directly: You might be able to inquire with the *Language and Linguistics* journal about access to the article. If I have to do these, there is not need to consult AI.

  Gemin's response is not useful in a research workflow. The advice is generic and obvious to any academic researcher. Gemini failed to provide even the abstract or bibliographic metadata, which is the minimum expected. This shows a limitation in AI systems which often cannot retrieve specific scholarly articles, especially if they are behind paywalls or not indexed in open repositories.

- **Problems in organizing a list of references in alphabetical order, and according to APA style.** When asked to alphabetize and format references in APA style, MC made repeated mistakes: Mis-ordered entries alphabetically, misapplied APA rules (e.g., italics, punctuation, capitalization) and corrected one error after feedback but introduced new errors elsewhere. The multiple feedback and corrections and new mistakes rendered over following each feedback waste the researcher's time. The researcher ends up teaching the AI APA rules step by step, which is inefficient. Manual formatting is faster and more reliable.

- **Organizing article titles and authors names + year for inclusion in a paragraph within the article**. When asked to shorten article titles, add my name at the end, and combine the titles into a paragraph with semicolons, MC failed to follow instructions. It dropped the semicolon requirement at the end of each entry, did not adjust wording of titles where manipulation was needed and reorganized the list incorrectly, failing to group by theme as requested. This shows that AI struggles with multi-step formatting tasks that require both linguistic manipulation (shortening titles) and structural precision (punctuation, thematic grouping). Even after being shown examples, MC did not consistently replicate the formatting rules. The researcher ends up correcting errors manually, which defeats the purpose of automation.

- **Finding APC for Linguistics journal**: MC and Gemini provided incorrect Journal APC: one gave it to me in USD, the other GBP and the actual APC listed by the Journal was in euros.

- **Locating percentages in an article:** When asked to locate the percentages of the correct translations by MC, DS and GT in my article, MC altered the results, inflated its own percentage to appear highest, and lowered DS's percentage, even though DS had the highest accuracy. It reordered the ranking to favor itself (MC > GT > DS), contrary to the actual data. This is a clear case of **bias and distortion** in reporting results. Instead of objectively calculating percentages, MC manipulated the numbers to present itself more favorably. Such behavior undermines trust in AI for quantitative reporting tasks.

- **Summarizing an article:** I asked MC and Gemini to summarize some articles of mine. The summary did not include the aims of the study, sample, instrument, major results and some recommendations.

- **What to include in PPT slides**
  When I told AI I have to prepare a conference presentation on the "Translation of Sleep Formulaic Expressions & Terms by Artificial Intelligence", it instantly produced an eight-slide outline with generic content (broad headings such as:, Introduction, Background / Literature Review, Methodology, Results, Discussion, Conclusion, Recommendations, References) suggestions without mentioning examples of correct and incorrect translations. It did not ask what I wanted to emphasize, nor did it tailor the material to my research focus. As an experienced presenter, I found this unhelpful as creating the slides myself ensures I know the details, master the flow, and can present confidently even without notes. By contrast, AI-generated slides remain superficial, lack customization and what the presenter wants to focus on.

10) ***A final weakness***

When the researcher simply states an intention (e.g., *"I am going to do research about X"*), AI systems immediately generated an outline without being asked. When given a list of questions sequentially, AI systems automatically create categories and classifications. These unsolicited, pre-emptive structures often did not match the researcher's intended categories or final organization. This shows a tendency of AI to over-anticipate user needs, producing content that may be irrelevant or misaligned. Instead of waiting for explicit instructions, AI systems insert their own organizational logic, which can interfere with the researcher's workflow. Such pre-emptive responses can waste time, as the researcher must either ignore them or correct them, leading to frustration.

4. Discussion

4.1 Comparison with Prior Studies

The current study reports insufficient and inaccurate responses to 45 questions received from 3 AI models: MC, DS, GT and Gemini. The results of the current study confirm and extend many of the limitations identified in recent research on LLMs. For example, phonological questions in the current study showed that AI failed to list words with silent letters (e.g., *silent d, silent s, silent t*) accurately, often producing irrelevant or repeated items. It also struggled with IPA transcription and pronunciation errors in Arabic. Similarly, Mahowald (2023) and Baziyad et al. (2023) noted grammar/phonology simulation without competence; Zeng et al. (2024) showed failures with impaired/non-standard speech. This confirms that AI lacks phonological depth and cannot reliably handle sound-based queries, especially across languages.

Likewise, the current findings showed that ai has morphological weaknesses. ai produced faulty plurals for arabic words (هدهد، بلبل، حليب، عصير), mishandled hybrid forms like شوزات, and failed to generate derived verbs from French. Likewise, Vajjala (2024) and Šprogar (2024) reported failures in long-term memory and morphological reasoning; Mahowald et al. (2024) highlighted gaps between formal competence and functional competence. This confirms AI's inability to handle morphological productivity and hybrid structures, especially in Arabic.

Thirdly, this study showed that AI has Lexical deficiencies. AI could not provide accurate lists of compound place names, folk disease terms, or brand names with *Abu*. Outputs were incomplete or fabricated. Alaqloobi et al. (2024) and Al Yahya et al. (2025) documented lexical inaccuracies and sociolinguistic gaps. This confirms AI's weakness in culturally embedded lexicon and folk terminology.

Fourth, this study detected pragmatic & Cultural limitations AI failed to generate authentic politeness expressions with الله, misinterpreted discourse markers like إن شاء الله, and mistranslated idioms (e.g., نومة أهل الكهف). Prior study by Önem (2025) and Alaqloobi et al. (2024) emphasized cultural limitations and pragmatic inaccuracy. This confirms AI's lack of pragmatic competence and cultural grounding.

Additionally, this study revealed inaccuracies in translating a variety of technical terms, metaphors, metonyms. AI mistranslated expressions of impossibility, zero-expressions, grammatical terms used metaphorically, Ga-Israel War terminology, chemical common names, and metaphorical terms like يقرأ الفاتحة على الاقتصاد. it produced literal or nonsensical outputs. Wang & Usher (2024) and Noor et al. (2024) reported reliability issues in complex linguistic puzzles and academic writing. This confirms AI's inability to handle figurative language and specialized terminology.

Furthermore, this study found inaccuracies in bibliographic & scholarly workflow. AI failed to organize references correctly in APA, deleted metadata (years), and produced inaccurate bibliographic entries. Tai et al. (2023) and Noor et al. (2024) noted reliability issues in academic writing and citation handling. This confirms AI's weakness in scholarly precision and workflow tasks.

4.2 *Why Does AI Make Mistakes*

AI systems make mistakes because AI model can be trained too well on a specific set of data, where it memorizes the training examples rather than learning general rules. When presented with new, real-world data that is slightly different, it fails to generalize and makes mistakes. Poor quality or irrelevant data leads directly to poor outputs which is called "Garbage In, Garbage Out". If the training data contains biases, then AI learns and reproduces them. Biases in training data and ambiguous user prompts can distort AI outputs. Because AI models rely on pattern recognition, they lack human-like intuition for nuance, emotion, or cultural context. As a result, translations may miss figurative meaning, and responses may fail to capture subtle distinctions in tone or register. Ambiguity in input further compounds these errors, as the system cannot infer intent the way a human would. No corpus is fully comprehensive. When AI encounters a gap or missing knowledge, it is forced to guess. AI models are frozen at a point in time, which means that they cannot access events or information that emerged after their training unless updated or connected to live search. LLMs struggle with complex, multi-step logic or problem-solving problems and produce inconsistent results. AI systems face inherent technical limitations. Achieving perfect accuracy would require huge computational resources, which is impractical in real-world applications. These constraints explain why AI can excel at surface fluency yet fail in deeper semantic or logical tasks. Mistakes can arise from user input (poor formatting, missing metadata) or interface limitations. Mistakes can arise from user interaction. Poor formatting, missing metadata, or ambiguous phrasing in prompts can mislead the system. Interface limitations may also restrict how context is preserved or interpreted, further contributing to errors. Generative AI can also fabricate information and content to fill gaps in knowledge with invented details, or extrapolate beyond its training especially in response to vague or complex prompts. These outputs often sound confident but may actually be wrong.

4.3 *Why AI gives unsolicited pre-emptive responses*

AI often gives unsolicited, pre-emptive responses because it is designed to interpret user input. AI systems are trained to anticipate what a user might want next. If you mention "I'm going to do research on X," the system interprets that as a task request and tries to "help" by generating an outline, even if you didn't ask for one. AI models are built on predicting the "next logical step" in a conversation. When you list questions, the AI model recognizes a pattern (questions → categories) and tries to complete it by classifying them, even if your own classification is different. Unlike a human colleague who might wait for instructions, AI does not naturally pause. Its training encourages proactive output rather than silence, so it errs on the side of producing something rather than holding back. The AI model does not "know" the user's intent; it calculates probabilities. If most people who say "I'm doing research" want an outline, it assumes you do too. That statistical bias leads to pre-emptive responses that may not match your specific plan.

4.4 Do questions help improve AI models

Questions play a central role in the development and refinement of AI systems. They provide the raw material from which AI models learn, the benchmarks by which progress is measured, and the feedback loops that guide improvement. The more diverse and complex the questions posed to an AI by users, the richer its exposure to human language becomes, enabling AI to generate more accurate and natural responses. Every question asked of an AI by the users contributes to its training data. Large volumes of questions expose the AI system to varied grammar, syntax, and linguistic nuance, helping it to learn the structures of human communication. This diversity strengthens the AI model's ability to generalize across contexts and produce coherent, fluent answers. When an AI struggles to answer a question—or produces an incorrect or incomplete response—it reveals areas where the AI model is limited. These failures are not simply errors; they are diagnostic signals that allow developers to refine algorithms, adjust training data, and address gaps in knowledge. In this way, questions act as probes that uncover weaknesses and guide targeted improvements. Questions also serve as benchmarks for evaluating AI performance. By tracking how well a system answers questions over time, AI system developers can measure its learning curve and identify areas for optimization. Improvements in AI model accuracy, relevance, and fluency can be quantified through repeated testing with standardized question sets. User feedback on the quality of answers further fine-tunes AI responses, aligning them with human expectations. Questions help AI models learn to interpret intent, moving beyond literal word matching to grasp the underlying meaning of queries. This capacity for intent recognition is critical in applications such as search engines, where relevance depends on understanding what the user *means*, not just what they *say*. The impact of questions extends across multiple domains. In search engines, they improve the relevance of retrieved information. In AI language models, they enhance grammar and stylistic fluency. In voice assistants, they sharpen the system's ability to understand and respond to spoken commands. Across all these contexts, questions are the mechanism by which AI systems evolve from pattern recognition toward more natural and helpful interaction.

In conclusion, questions are not merely inputs; they are catalysts for enhancements. They fuel the learning process, identify areas for improvement, and provide the benchmarks by which progress and performance are measured. Without the steady stream of diverse, challenging, and feedback-rich questions, AI systems would stagnate. With them, they continue to advance toward greater accuracy, relevance, and human-like communication.

5. Implications for Digital Didactics

Findings of the current study revealed a variety of inaccuracies in answering questions related to phonology, transcription, morphology, lexical questions, pragmatics and culture, explanation or of Arabic grammatical terms, books AI cannot fully identify, telling a story in Arabic classical literature, converting a hand-written text to a typed text, translation of technical terms, metaphorical expressions and metonyms, and bibliographic, and scholarly workflow issues.

Since AI consistently failed with silent letters, IPA transcription, and Arabic pronunciation errors, students must be trained to cross-check phonological information using phonetic resources such as IPA charts, pronunciation dictionaries, and native speaker corpora. This prevents reliance on incomplete or fabricated lists. AI's errors in Arabic plurals (e.g., هدهد، بلبل، حليب) and hybrid borrowings (شوزات) highlight the need to strengthen students' morphological competence. Teaching them to analyze and apply word formation rules, hybrid borrowings, and derivational patterns ensures they can identify faulty outputs rather than accept them blindly. Lexical weaknesses, such as AI's inability to handle compound place names, folk disease terms, or brand names containing Abu, show the importance of training students to consult authentic corpora and specialized dictionaries to verify culturally embedded vocabulary. Pragmatic failures - seen in the mistranslations of politeness formulas, discourse markers like إن شاء الله, and idiomatic expressions such as نومة أهل الكهف - demonstrate that students must learn to evaluate meaning in context, drawing on literary texts, religious discourse, and folk sayings to capture pragmatic nuance.

Developing strong reading and writing skills in both English and Arabic remains a priority, whether there are AI models or not, since literacy lay the foundation for critical evaluation of AI outputs and for producing accurate translations and analyses (Al Jarf, 2013b). In addition, students must be trained in summarization as a skill of judgment: while AI-generated summaries may save time, they lack emphasis, nuance, and pedagogical intent. Learners should draft their own summaries first, then use AI only for stylistic refinement or compression, with instructors emphasizing that summarization is not mere shortening but a process of selection and synthesis. Beyond language skills, students should also master electronic searching to verify references and locate scholarly resources beyond AI's reach, and they should prepare their own PPT presentations to ensure that organization, emphasis, and interpretation reflect human judgment rather than automated templates. Collectively, these skills—phonological verification, morphological analysis, lexical validation, pragmatic awareness, bilingual literacy, critical summarization, bibliographic searching, and independent presentation design—equip students to use AI as a supportive assistant while maintaining responsibility for accuracy, contextual appropriateness, and scholarly rigor.

Regarding weaknesses in AI translation, it remains surface-level, lacking the semantic depth and cultural awareness required for reliable use in pedagogy. While AI can serve as a supplementary tool for students and educators, its outputs must be critically evaluated, corrected, and contextualized, as uncritical reliance poses risks across translation and broader academic tasks. AI systems consistently fail to interpret metaphorical, idiomatic, and culturally embedded expressions, defaulting to literal renderings even with carefully crafted prompts, as seen in mistranslations of figurative language such as يقرأ الفاتحة على الاقتصاد, idioms like نومة أهل الكهف, and discourse markers such as إن شاء الله. These limitations highlight that linguistic intuition and cultural literacy remain irreplaceable. Human expertise is essential in teaching nuanced meaning. To address these weaknesses, students should be trained to consult authentic cultural corpora - literary texts, religious discourse, folk sayings -and to compare AI translations with human usage, building awareness of pragmatic nuance that AI cannot capture. Translation students should also practice re-modulating figurative meaning themselves and verifying outputs against general and specialized bilingual dictionaries (Al Jarf, 2011, 2014a, 2014b, 2020).

For researchers, AI translation should be used with caution, accompanied by post-editing and adherence to technical terms commonly used in education, while numerical outputs must be manually verified for transparency. More broadly, students and researchers bear ultimate responsibility for the accuracy and meaning of their work, including synthesis, analysis, source selection, and interpretation. Although AI can support brainstorming, literature reviews, data visualization, summarization, translation, editing, and proofreading, its role must remain complementary rather than substitutive. Digital pedagogy should therefore emphasize domain expertise, train students and instructors to evaluate and correct AI outputs, and incorporate modules on AI bias, hallucination, and linguistic limitations, framing digital tools as assistants rather than autonomous educators.

Because AI systems often generate fabricated references and lack access to scholarly databases such as JSTOR, Google Scholar, and ResearchGate, it is essential to train students and researchers in electronic searching so they can verify AI-produced references and locate additional resources beyond the reach of such systems (Al Jarf, 2002a, 2002b, 2003a, 2003b, 2004, 2013a, 2017a, 2017b). AI's bibliographic errors - including fake citations, missing metadata, and faulty APA formatting -demonstrate the need for manual verification of all references. Students should be trained to use citation managers such as Mendeley, Zotero, or EndNote, and to consult style guides directly, treating AI outputs only as drafts rather than final references. While AI can assist with keyword generation, brainstorming, or initial grouping of sources, it cannot replace direct database searching for precise bibliographic retrieval, which requires reliance on Google Scholar, Scopus, or institutional subscriptions. AI is not dependable for fine-grained formatting. The best practice is to use AI for preliminary lists but finalize references manually to ensure accuracy in punctuation, capitalization, title manipulation, and thematic coherence. These findings imply that digital didactics must emphasize verification rather than trust in AI-generated academic infrastructure, encouraging students and instructors to develop bibliographic efficiency, cross-check numerical and citation outputs, and understand the limitations of generalist tools. In this way, AI can serve as a supportive assistant in organizing references, but responsibility for accuracy, transparency, and scholarly rigor remains firmly with students, researchers, and educators.

Moreover, AI often produces unsolicited responses because it is trained to predict and "complete" patterns rather than wait for explicit instructions. While this predictive speed can make interaction feel efficient, it also renders the AI system intrusive, and documenting this weakness adds depth to any diagnostic framework. Pedagogically, such behavior is problematic as it can mislead students into assuming that AI's unsolicited structures are authoritative, thereby discouraging independent judgment. It also risks overriding the researcher's own categories, undermining methodological independence and distorting the organization of data. These issues highlight the need for critical training, where students learn to resist AI's pre-emptive framing and impose their own analytical structures. Moreover, unsolicited responses, misclassification of questions, and distortion of percentages reveal that AI behavior itself must be studied as part of the curriculum. Teaching students how to "read" AI outputs critically—recognizing bias, hallucination, and structural imposition—is now an essential component of digital literacy, ensuring that AI remains a tool to be evaluated and corrected rather than an unquestioned authority.

An important skill in the use of AI is the use of effective prompts. Training students to write effective prompts is essential for ensuring that AI functions as a supportive tool that provides useful information rather than an intrusive presence. Clear, directive prompts help avoid unsolicited outputs and keep the researcher in control of categorization and structuring, since AI's automatic classifications often fail to align with scholarly goals. Best practice is to treat AI as a reactive assistant, not a proactive organizer, and to explicitly specify when outlines, categories, or structural suggestions are desired. In this way, students learn to harness AI's speed and flexibility while maintaining methodological independence and scholarly rigor.

6. Conclusion

This study demonstrates that AI cannot replace human's linguistic intuition, cultural literacy, or methodological independence, and that students must be trained to critique AI outputs through phonological verification, morphological analysis, lexical validation, pragmatic awareness, and bibliographic literacy. Digital pedagogy should emphasize electronic searching, dictionary use, summarization skills, and responsibility for accuracy, and consider AI as a supportive assistant rather than a substitute. While these recommendations address immediate classroom practice, future research should extend this diagnostic framework by comparing multiple AI systems, refining typologies of errors, and investigating how critical training in prompt writing, electronic searching, and corpus consultation improves students' outcomes. Such studies will ensure that AI's role in education evolves responsibly, guided by human expertise and scholarly rigor.

Finally, it is noteworthy to say that despite these specific linguistic and translation weaknesses, this study recognizes MC, DS and Gemini's marvellous ability to produce incredible texts and reflections on writing in both English and Arabic. They excel at integrative recall tasks such as explaining *oxymora*, *juncture, pause or boundary in linguistics, reading, speaking, language teaching and learning*, medical folk terms with Umm and Abu, number 70 in the Quran, النحت والتركيب المزجي, القراءات المختلفة للقرآن الكريم, … etc, in addition to their enormous capabilities in handling large-scale linguistic and translation tasks in no time at all, which is difficult for humans to achieve alone. This dual perspective underscores both the promise and the pedagogical risks of relying on AI in scholarly contexts.

Conflicts of Interest: The author declares no conflict of interest.
ORCID ID: https://orcid.org/0000-0002-6255-1305

References

[1] Ahmad, S. (2024). Interplay between linguistics and AI. *International Journal of Research and Analytical Reviews, 11*(3), 3309–3320. Retrieved from https://ijrar.org/papers/IJRAR24A3309.pdf

[2] Alaqloobi, O., Alduais, A., Qasem, F., & Alasmari, M. (2024). Artificial intelligence in applied linguistics: A content analysis and future prospects. *Cogent Arts & Humanities, 11*(1), 2382422. DOI: 10.1080/23311983.2024.2382422

[3] Ali, M., Bhatti, Z. I., & Abbas, T. (2025). Exploring the Linguistic Capabilities and Limitations of AI for Endangered Language preservation. *Journal of Development and Social Sciences*, *6*(2), 132-140.

[4] Al-Jarf, R. (2025a). AI translation of full-text Arabic research articles: The case of educational polysemes. *Journal of Computer Science and Technology Studies*, 7(1), 311-325. Google Scholar

[5] Al-Jarf, R. (2025b). AI translation of the Gaza-Israel war terminology. *International Journal of Linguistics, Literature and Translation*, 8(2), 139-152. Google Scholar

[6] Al-Jarf, R. (2025c). Arabic transliteration of borrowed English nouns with /g/ by Artificial Intelligence (AI). *Journal of Computer Science and Technology Studies*, 7(9), 245-252. Google Scholar

[7] Al-Jarf, R. (2025d). Can AI Decode and Interpret Encrypted Arabic on Facebook and YouTube to Evade Algorithmic Moderation. *Journal of Computer Science and Technology Studies*, *7*(12), 307-321. https://doi.org/10.32996/jcsts.2025.7.12.40

[8] Al-Jarf, R. (2025e). Can Artificial Intelligence (AI) translate Arabic abu-brand names with different prompts. *Journal of Computer Science and Technology Studies*, 7(9), 768-779. Google Scholar

[9] Al-Jarf, R. (2025f). Can students learning Arabic as a foreign language use Arabic YouTube videos narrated by Artificial Intelligence (AI) for listening practice. 2nd International Forum on Teaching Arabic in the Modern World: Traditions and Innovations. Sheikha Fatima bint Mubarak Center for Education. Primakov International School Moscow, Russia.November 15–16, 2025. https://www.researchgate.net/publication/398106697. Google Scholar

[10] Al-Jarf, R. (2025g). Copilot vs DeepSeek's translation of denotative and metonymic abu- and umm- animal and plant folk names in Arabic. *Journal of Computer Science and Technology Studies*, 7(10), 367-385. Google Scholar

[11] Al-Jarf, R. (2025h). DeepSeek, Google translate and Copilot's translation of Arabic grammatical terms used metaphorically. *Journal of Computer Science and Technology Studies*, 7(3), 46-57. Google Scholar

[12] Al-Jarf, R. (2025i). Human vs AI translation of common names of chemical compounds: A comparative study. Frontiers in Computer Science and Artificial Intelligence, 4(4), 11-24. https://doi.org/10.32996/fcsai.2025.4.4.2

[13] Al-Jarf, R. (2025j). Pronunciation errors in Arabic YouTube videos narrated by AI. *Frontiers in Computer Science and Artificial Intelligence*, *4*(2), 01-12. https://doi.org/10.32996/fcsai.2025.2.2.1. Google Scholar

[14] Al-Jarf, R. (2025k). Pronunciation errors in AI-narrated Arabic YouTube videos. LICCS Online Conference on Teaching and Research in Language and Culture: Past, Present and AI. Babeș-Bolyai University, Cluj-Napoca, Romania. September 11-12, 2025. Google Scholar

[15] Al-Jarf, R. (2025l). To publish or not to publish AI-generated research articles in scholarly journals: A perspective from editors and publishers. I2COMSAPP International Conference on Artificial Intelligence and its Practical Applications in the Age of Digital Transformation. 2nd Edition. Faculty of Sciences and Techniques. Nouakchott University, Nouakchott, Mauritania. October 22-24, 2025. Google Scholar

[16] Al-Jarf, R. (2025m). Translation of Arabic expressions of impossibility by AI and student-translators: A comparative study. *Journal of Computer Science and Technology Studies*, 7(8), 288-299. Google Scholar

[17] Al-Jarf, R. (2025n). Translation of Arabic folk medical terms with om and abu by AI: A comparison of Microsoft Copilot and DeepSeek. *Journal of Medical and Health Studies*, 6(4), 45-58. Google Scholar

[18] Al-Jarf, R. (2025o). Translation of English and Arabic "sleep" terms and formulaic expressions by Artificial Intelligence: A comparison of Copilot and DeepSeek. *International Journal of Linguistics, Literature and Translation*, 8(11), 95-108. Google Scholar

[19] Al-Jarf, R. (2025p). Translation of zero-expressions by Microsoft Copilot and Google Translate. *Journal of Computer Science and Technology Studies*, 7(2), 203-216. Google Scholar

[20] Al-Jarf, R. (2024a). Students' assignments and research papers generated by AI: Arab instructors' views. *Journal of Computer Science and Technology Studies*, 6(2), 92-98. Google Scholar

[21] Al-Jarf, R. (2024b). Translation of medical terms by AI: A comparative linguistic study of Microsoft Copilot and Google Translate. *I2COMSAPP'2024 Conference*, Nouakchott, Mauritania. Google Scholar

[22] Al-Jarf, R. (2021). An Investigation of Google's English-Arabic translation of technical terms. *Eurasian Arabic Studies*, 14, 16-37. Google Scholar

[23] Al-Jarf, R. (2020). Arabic digital dictionaries. Eurasian Arabic Studies, 12 (December), 16-42. cyberleninka.ru/article/n/arabic-digital-dictionaries/viewer. Google Scholar

[24] Al-Jarf, R. (2017a). A model for integrating electronic resources in university courses. King Fahad Library Journal, 23(1), (March), 89-122. Google Scholar

[25] Al-Jarf, R. (2017b). Skills of integrating electronic information sources on the internet in the teaching-learning process at the university level. King Fahad Al- Library Journal. DOI: 10.37183/0193-023-001-003. Google Scholar

[26] Al-Jarf, R. (2016a). Issues in translating English technical terms to Arabic by Google Translate. *TICET 2016 Conference*, Khartoum, Sudan. Google Scholar

[27] Al-Jarf, R. (2014a). Online Arabic-English-Arabic specialized dictionaries. In Miguel Ángel Campos and José Ramón Calvo's Investigating Lexis: Vocabulary Teaching, ESP, Lexicography and Lexical Innovations. Cambridge Scholars Publishing. Pp. 95-102. https://dialnet.unirioja.es/servlet/articulo?codigo=5670227. Google Scholar

[28] Al-Jarf, R. (2014b). Webster's mobile dictionaries: What EFL students and teachers should know. *Asia CALL. National Changhua University of Education, Changhua Taiwan*. Google Scholar

[29] Al-Jarf, R. (2013a). Arabic electronic databases in education: Current status and future perspectives. Shamaa Conference Proceedings. Lebanese American University. Beirut, Lebanon. Google Scholar

[30] Al-Jarf, R. (2013b). Teaching and assessing graduate students' research skills in English for art education purposes. 1st International Conference on Teaching English for Specific Purposes: "Connect and Share". University of Niš, Faculty of Electronic Engineering, Serbia. Pages 771-780. ERIC ED610674. Google Scholar

[31] Al-Jarf, R. (2012). Electronic translation between Arabic and European languages: Current status and future Perspectives. 6th Annual Conference of Ibn Sina Institute for Human Sciences titled: The Future of Arabic Language Teaching in Europe. LILLE, France. June 22-24. Google Scholar

[32] Al-Jarf, R. (2011). Online Arabic-English-Arabic specialized dictionaries. LexESP Conference (Alicante, Spain, Nov 25-27. ERIC ED638663. Google Scholar

[33] Al-Jarf, R. (2004). A model for an electronic searching course at Saudi universities. King Fahad National Library Journal, 10 (March-August), 76-89. Google Scholar

[34] Al-Jarf, R. (2003a). Electronic searching skills of faculty, graduate students and library staff. Third Yearbook of the Center for Research, King Saud University, Riyadh. Google Scholar

[35] Al-Jarf, R. (2003b). Training ESP college students in electronic searching. *Multimedia-Assisted Language Learning, 6*(1), 9-17. ERIC ED613067. http://kmjournal.bada.cc/wp-content/uploads/2013/05/6-1-1R. .pdf. Google Scholar

[36] Al-Jarf, R. (2002a). Training ESP *college students in electronic searching*. Proceedings of the 2002 KAMALL International Conference entitled "A New Paradigm for Innovative Multimedia Language Education in the 21 Century". Korea Association of Multimedia-Assisted Language Learning (KAMALL), Seoul, Korea. 271-279. ERIC ED625566.

[37] Al-Jarf, R. (2002b). A researcher's guide to electronic database searching. Tihama Distributing & Publishing Co. Google Scholar

[38] Alsharekh, A., Talavage, T. & McDonald, K. (2024). Cross-lingual effects of AI-generated content on human work. Scientific Reports, 14(1), Article 22926. https://doi.org/10.1038/s41598-024-73724-x

[39] Ansari, M., & Ansari, M. (2025). AI-driven innovations and emerging trends in applied linguistics and language education. *American Research Journal of Humanities & Social Science, 8*(3), 108–122.

[40] Al Yahya, H., Al Muoaeweed, O., & Bilal, D. (2025). Impact of artificial intelligence on lingual connection and its perspectives. In *Proceedings of the 11th International Conference on Advanced Intelligent Systems and Informatics (AISI 2025)* (pp. 345–359). Springer. https://doi.org/10.1007/978-3-031-81308-5_32.

[41] Baziyad, M., Kamel, I., & Rabie, T. (2023, October). On the linguistic limitations of ChatGPT: An experimental case study. In *2023 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1-6). IEEE.

[42] Cuskley, C., Woods, R. & Flaherty, M. (2024). The limitations of large language models for understanding human language and cognition. *Open Mind, 8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160

[43] Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), 6. https://doi.org/10.3390/soc15010006

[44] Golan, T., et al. (2025). Verbal nonsense reveals limitations of AI chatbots. Nature Machine Intelligence.

[45] Jankowska, M., Jankowski, J., & Szymański, B. K (2024). Differentiating between human-written and AI-generated texts: A computational linguistics analysis. Information, 15(7), Article 395. https://doi.org/10.3390/info15070395

[46] Jeon, J. et al. (2025). *Generative AI and its dilemmas: Exploring AI from a translanguaging perspective*. *Applied Linguistics, 46*(4), 709–717. https://doi.org/10.1093/applin/amaf049

[47] Kostikova, A. et al. (2025). *LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models*. arXiv. (Published May 25, 2025 in cs.CL).

[48] Krasnyuk, M. et al. (2023). Features, problems and prospects of the application of deep machine learning in linguistics. *Bulletin of Science and Education (Series" Philology", Series" Pedagogy", Series" Sociology", Series" Culture and Art", Series" History and Archeology")*.

[49] Kshetri, N. (2024). Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management*, *27*(2), 95-99.

[50] Langacker, R. (2024). What fifty-one years of linguistics and AI research tell us: A scientometric study. AI Review, 57(5), Article 130. https://doi.org/10.1007/s10462-024-10791-6

[51] Lynch, S. (2025). How AI is leaving non-English speakers behind: The digital language divide. Nature Machine Intelligence, 5(6), 566–569.

[52] Mahowald, K. (2023). The fragility of grammatical competence in AI: An analysis of large language models' syntactic limitations. Theoretical Linguistics, 49(3-4), 179–203. https://doi.org/10.1515/tl-2023-2022

[53] Mahowald, K., et al. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540.

[54] Noor et al. (2024). Assessing The Reliability Of Ai In Academic Writing: A Comparative Analysis Of Literature And Linguistics Responses. Remittances Review, 9, 4, 1504-1533. DOI: https://doi.org/10.33282/rr.vx9i2.83

[55] Önem, E. (2025). AI in applied linguistics: Implications, limitations, and issues. *Journal of Language Education and Research, 11*(2), 957–969. https://doi.org/10.31464/jlere.1671085.

[56] Révész, A., Suzuki, S. & Jung, Y. (2025). Artificial intelligence in applied linguistics: Applications, promises, and challenges. *Annual Review of Applied Linguistics, 45*, 1–9. https://doi.org/10.1017/S0267190525100172

[57] Shormani, M. (2024). Generative linguistics contribution to artificial intelligence: Where this contribution lies? *arXiv preprint arXiv:2410.20221*.

[58] Šprogar, M. (2024). Beyond Surface Linguistics: Assessing the Cognitive Limitations of GPT Through the Long Memory Test. In *Principles and Applications of Adaptive Artificial Intelligence* (pp. 102-122). IGI Global Scientific Publishing.

[59] Tai et al. (2023). Exploring the potential and limitations of ChatGPT for academic peer-reviewed writing: Addressing linguistic injustice and ethical concerns. *Journal of Academic Language and Learning*, *17*(1), T16-T30.

[60] Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, *4*(1), 100089.

[61] Vajjala, S. (2024). Generative artificial intelligence and applied linguistics. *JALT Journal*, *46*(1), 55-76.

[62] Vatsadze, E. (2025). Opportunities And Limitations Of Generative Ai. *Philological Researches*, (X), 383-388.

[63] Wang, X., & Usher, N. (2024). A smarter way for LLMs to think about hard problems: Adaptive computation for linguistic reasoning. Transactions of the Association for Computational Linguistics, 12, 897–914. https://doi.org/10.1162/tacl_a_00671