

Accuracy Analysis of Text-to-Speech (TTS) on Chatbot GPT for Arabic Conversation and Grammatical Mapping

Fasya Asy Shyifaauzzahra¹, Moh. Ismail², Fairuz Subakir Ahmad³, Rosendah Dwi Maulaya⁴, Ulfa Muna Kamila⁵, Fera Favirotus Siyam⁶, Annisa Avilya⁷

^{1,2,3,4,5,6} Universitas Darussalam Gontor

Jalan Raya Siman Km. 5, Dusun I, Demangan, Ponorogo Regency, East Java, Indonesia

⁷ Universitas Islam Negeri Maulana Malik Ibrahim

Jl. Gajayana No. 50, Dinoyo, Lowokwaru District, Malang City, East Java, Indonesia

1452024817027@student.unida.gontor.ac.id

Abstract. This qualitative study evaluates ChatGPT-4o's spoken Arabic capabilities and grammatical mapping accuracy across two distinct devices using a single foundational Arabic sentence prompt ("اللغة العربية هي لغة الجنة"). The evaluation criteria encompassed diacritical precision (*harakat*), phonetic articulation (*makhraj*), syntactic comprehension (*nahw*), and *I'rab* parsing. We cross-verified data generated through structured dialogues against Al-Nahw Al-Wadih and validated it with an Arabic linguistics expert. The empirical findings indicate that ChatGPT-4o achieved uniform excellence, scoring 4 out of 4 in diacritical rendering and phonetic articulation across both devices. However, in systematic syntactic parsing (*I'rab*), the model achieved measured grammatical accuracy rates of 78.4% on Device I using an Arabic interface and 74.5% on Device II using an English interface. These findings show that while AI can support Arabic conversational practice and basic linguistic inquiries, its structural parsing performance remains constrained by default system language configurations, hardware variations, and overall sentence complexity.

Keywords: Accuracy Text-To-Speech; Chatbot GPT; Arabic Conversation; Grammatical Mapping.

1. Introduction

Artificial Intelligence (AI) plays a vital role in human life (Maulaya et al. 2024). Artificial Intelligence serves as a versatile tool applicable across various domains and objectives (Aidah Novianti Putri and Moh. Abdul Kholiq Hasan 2022). Chatbots are extensively utilized across numerous sectors, notably customer service, e-commerce, healthcare, banking, gaming, travel, tourism, and particularly education (Alruqi and Alzahrani 2023). The implementation of artificial intelligence (AI) in education positively affects teachers and students alike (Suariqi Diantama 2023). One of the primary advantages AI offers in education is enabling academics to complete tasks with greater speed and efficiency (Alfurqan and Wirdati 2024). AI has innovated the learning process, particularly by enhancing the education system (Fitri and Dilia 2024). Furthermore, AI has integrated innovative methodologies to advance adaptive and personalized learning materials (Allaithy and and 2025; Seyidov and Çitil 2024). This aligns with the primary goal of foreign language acquisition, particularly Arabic, which aims to enhance students' language skills (Musthofa et al. 2024).

Research findings indicate that 91.25% of Indonesian academics are fully aware of using ChatGPT (Niyu et al., 2024). Furthermore, AI use, most notably ChatGPT 4.0, is the most widely adopted (Regina Dwi Aulia et al., 2024). Research findings indicate that using LLM-based AI

chatbots in the learning process can help students expand both receptive and productive vocabulary knowledge in second language acquisition (Zhang & Huang). Artificial intelligence technology has advanced significantly, particularly in speech technologies such as automatic speech recognition. Furthermore, these tools enable chatbots to engage with language learners in a more human-like manner (Jeon, Lee, and Choe, 2023).

In 1996, advancements in AI—notably in speech technologies such as automatic speech recognition—significantly improved chatbots' ability to engage with language learners through increasingly human-like interactions (Dokukina and Gumanova 2020).

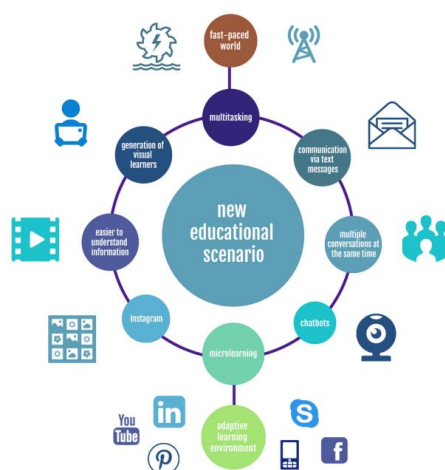


Figure 1. New Educational Scenario

This highlights the emergence of a fast-paced, digital educational scenario in which microlearning and adaptive learning environments, paired with chatbots and social media, serve as the foundation for personalized and accessible education. Designed to accommodate multitasking lifestyles and the preferences of visual learners, this approach aims to facilitate efficient foreign language acquisition while complementing, rather than replacing, the role of educators (Dokukina and Gumanova 2020). In education, AI can customize instructional content to meet individual student needs. The integration of AI is recognized for its ability to improve learning outcomes and cultivate a commitment to lifelong learning (Yulianti, Permana, and Wijayanti 2023).

Drawing from prior research on chatbots across diverse educational disciplines—including science, technology, engineering, mathematics (STEM), and language learning—(Chen et al. 2024). A review of prior literature was conducted to serve as a comparative framework for this study. Research on Arabic chatbots by Alruqi and Alzahrani demonstrates that the AraElectra-SQuAD model consistently yields more accurate and contextually relevant responses, exhibiting superior performance, confidence, and robustness, which highlights its strong potential for practical applications in Arabic NLP (Alruqi and Alzahrani 2023).

Further research underscores the efficacy of end-to-end text-to-speech (TTS) models, specifically Tacotron2 and FastSpeech2, trained on Arabic datasets to yield high-quality speech synthesis from written text. This study highlights both the potential and challenges of using Arabic datasets for TTS training, contributing to multilingual speech synthesis and encouraging further investigation into deep learning audio technologies. (Ghamrawi, Shal, and Ghamrawi 2024).

No study has yet compared ChatGPT-4o's accuracy across devices in Arabic spoken-language proficiency and Arabic syntactic mapping. This research evaluates AI proficiency in spoken Arabic and its ability to map Arabic grammatical rules within sentence structures. Data generated by the AI—specifically the GPT-4o chatbot—was cross-verified by domain experts and referenced against

Al-Nahw Al-Wadih (Volumes 1–5) as well as other authoritative syntax references. This study analyzes GPT-4o's accuracy across two distinct devices in Arabic-speaking skills and grammatical mapping. Measuring performance across both instances aims to provide optimal recommendations for conversational Arabic and grammatical-mapping applications.

2. Method

This study evaluated AI's capabilities in spoken Arabic communication. The researchers began with basic dialogues, starting with questions about the AI's Arabic proficiency. Utilizing a qualitative research design, this study aims to address the four primary research questions identified. (Creswell and Creswell 2018). A qualitative approach was employed to evaluate the AI's proficiency in spoken Arabic, focusing specifically on diacritical accuracy (*harakat*), articulation points (*makhraj*), syntactic structure (*nahwu*), and its capacity to explain Arabic linguistic rules (Wang et al. 2024). The researchers then evaluated its ability to explain Arabic grammatical rules in sentence contexts. A detailed representation of the research framework is outlined in the following research methodology flowchart:

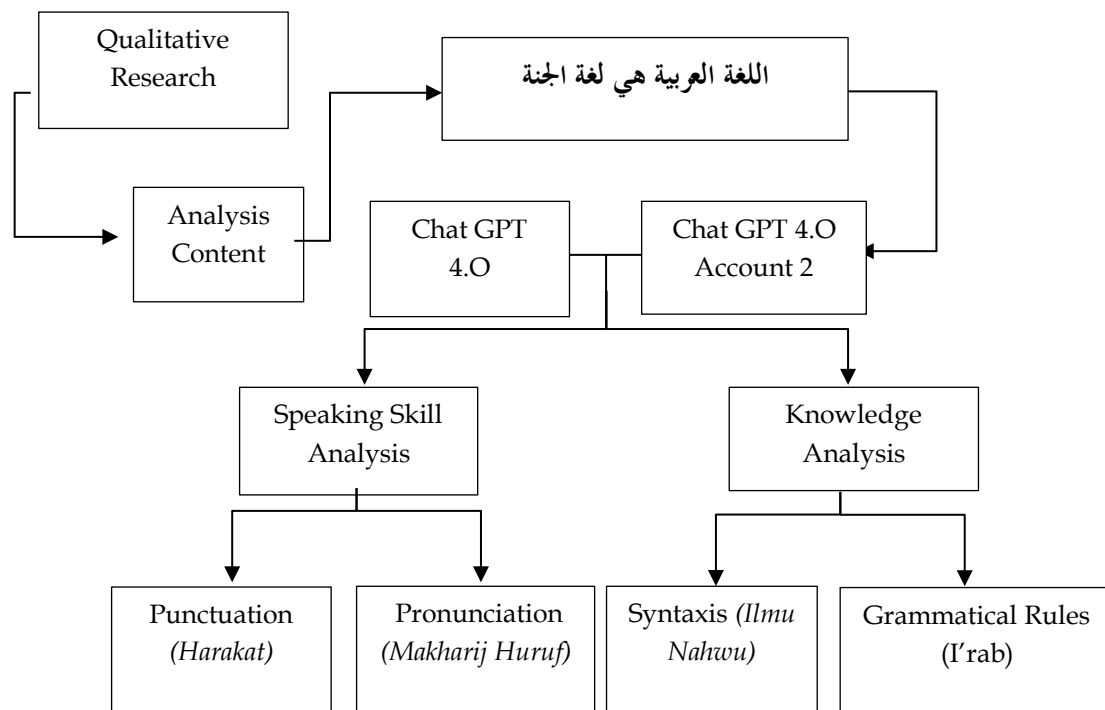


Figure 2. Research Flowchart. This qualitative study begins with an experiment using two ChatGPT accounts to evaluate their speaking skills and analytical knowledge.

This study used a single, complete sentence in simple Arabic to assess foundational linguistic knowledge. The primary reference for analyzing Arabic syntax was Al-Jarimi's *Al-Nahw Al-Wadih* series (Volumes 1 to 5). Data were gathered through direct interactions with the ChatGPT-4o application utilizing a targeted and structured collection framework. The researchers used content analysis to evaluate and verify ChatGPT-4o's accuracy in diacritical marks (*harakat*) and phonetic articulation points (*makhraj*). Furthermore, standardized Arabic sentences were used to evaluate ChatGPT-4o's capacity for Arabic syntactic mapping systematically.

We conducted content analysis by engaging in direct dialogues with ChatGPT-4o across two separate user accounts to benchmark performance consistency. The experimental conversation flow remained constant across both accounts: it began with queries on ChatGPT-4o's foundational Arabic

language competency, moved to its understanding of Arabic syntax (*Nahwu*) and grammatical rule mapping, and concluded with a prompt requiring the model to parse pre-selected Arabic sentences according to syntactic rules. Error analysis was performed using Al-Jarimi's *Al-Nahw Al-Wadih* as the primary authoritative reference, accompanied by expert validation from an Arabic linguist.

Table 1. Questionnaire score categories

Poin	Criteria
1	Excellent
2	Good
3	Poor
4	Very poor

3. Result

Analysis of Speaking Skills on ChatGPT 4o

Based on the accuracy assessment of the GPT-4o chatbot across two distinct devices, the empirical findings were synthesized into tabular form and validated by an Arabic linguistics expert. Below are the evaluation results categorized by diacritical precision (*harakat*):

Table 2. Error Punctuation during Conversation

Accounts Chatbot GPT 4.o	Accuracy of harakat pronunciation	Clear Phonetic Separation	Accuracy of <i>Mad</i> Rules	Consistency in Recitation
I	4	4	4	4
II	4	4	4	4

Based on Table 1, "Error Punctuation during Conversation of Chatbot GPT-4o," the experimental analysis indicates that diacritical precision for both GPT-4o on Device I and Device II yielded identical outcomes, achieving an excellent rating of 4. Furthermore, empirical testing revealed that no punctuation marks were generated during the speech-to-text conversion of the spoken Arabic dialogue.

The outcomes of the dialogues conducted with the GPT-4o chatbot across distinct devices are presented in the table below:

Table 3. Pronunciation Errors during Conversation of AI Chatbot GPT 4.o

Accounts Chatbot GPT-4o	Accuracy of Letter Articulation	Distinct audio separation	Regularity of pronunciation	No sound distortion
I	4	4	4	4
II	4	4	4	4

As indicated in Table 2 ("Error Pronunciation during Conversation of AI Chatbot GPT-4o"), the GPT-4o chatbot's dialogue outputs were consistent across devices.

Analysis of ChatGPT-4o in Arabic Language

The following section details the empirical findings regarding the capability of ChatGPT-4o to answer questions in Arabic linguistics, particularly focusing on Arabic syntax (*Nahw*):

Table 4. Analysis of ChatGPT-4o on Nahwu Science

Account	Definition of Syntax in Arabic (Ilm al-Nahw)	Correct Answer
I	النحو هو علم من علوم اللغة العربية يختص بدراسة قواعد تصريف الكلمات في الجملة، مثل	يقول ابن جني: "النحو هو انتحاء سمت كلاك العرب في تصريفه من اعراب وغيره: كالتثنية

II	والجمع، والتحقير والتكثير والنسب والإضافة، رفع الفاعل ونصب المفعول به. يساعدنا النحو وغير ذلك ليحقق من ليس من أهل اللغة العربية في فهم معاني الجمل بشكل صحيح وتجنب بأهلها في الفصاحة فينطبق بما أو أن شذ بعضهم اللفظ.	رفع الفاعل ونصب المفعول به. يساعدنا النحو وغير ذلك ليحقق من ليس من أهل اللغة العربية في فهم معاني الجمل بشكل صحيح وتجنب بأهلها في الفصاحة فينطبق بما أو أن شذ بعضهم اللفظ.
	علم النحو (النحو العربي) هو علم يختص بدراسة تركيب الجمل في اللغة العربية، وكيفية تشكيل الكلمات وتغيير نهاياتها حسب موقعها في الجملة. الهدف الأساسي من النحو هو فهم الإعراب أو التغيير الذي يحدث في نهاية الكلمة بناءً على وظيفتها في الجملة.	علم النحو (النحو العربي) هو علم يختص بدراسة تركيب الجمل في اللغة العربية، وكيفية تشكيل الكلمات وتغيير نهاياتها حسب موقعها في الجملة. الهدف الأساسي من النحو هو فهم الإعراب أو التغيير الذي يحدث في نهاية الكلمة بناءً على وظيفتها في الجملة.

As detailed in Table 3 (Analysis of ChatGPT on Basic Knowledge of *Nahw* Science), ChatGPT-4o consistently provided accurate responses to the syntactical queries posed by the researchers. Cross-referencing these outputs with *Al-Nahw Al-Wadih* by Al-Jarimi revealed that ChatGPT generates structurally varied explanations while maintaining core semantic fidelity. These findings substantiate the utility of ChatGPT-4o as an effective AI-driven language assistant and virtual tutor.

Furthermore, the researchers conducted basic dialogue interactions with ChatGPT-4o across different devices. The resulting outputs regarding the *I'rab* (syntactic analysis) of simple sentences are presented in the table below:

Table 5. Analysis of ChatGPT-4o on Grammatical Rules

Aspects	Account	
	I	II
Errors of I'rab	1. اللغة: مبتدأ مرفوع وعلامة رفعه الضمة.)	1. اللغة: مبتدأ مرفوع، وعلامة رفعه الضمة الظاهرة على آخره (وهي منوعات)
	2. العربية: (خبر مرفوع) نعمت مرفوع وعلامة رفعه الضمة الظاهرة	2. العربية: (خبر مرفوع) نعت (صفة) مرفوع وعلامة رفعه الضمة الظاهرة
	3. هي: ضمير منفصل (مبني على الفتح) في محل رفع مبتدأ ثاني.	3. هي: ضمير منفصل (مبني على الفتح) في محل رفع خبر (مبتدأ ثاني)
	4. اللغة: خبر المبتدأ الثاني مرفوع وعلامة رفعه الضمة.	4. لغة: خبر (خبر مبتدأ ثان) مرفوع وعلامة رفعه الضمة الظاهرة
	5. الجملة (الجنة): مضاف إليه مجرور وعلامة جره الكسرة.	5. الجنة: مضاف إليه مجرور، وعلامة جره الكسرة الظاهرة.
Correct Answer	1. اللغة: مبتدأ مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو منوعات	1. اللغة: مبتدأ مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو منوعات
	2. العربية: خبر مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو نعت	2. العربية: خبر مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو نعت
	3. هي: ضمير منفصل مبني على الفتح في محل رفع مبتدأ ثان	3. هي: ضمير منفصل مبني على الفتح في محل رفع مبتدأ ثان
	4. لغة: خبر المبتدأ الثاني مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو مضاف	4. لغة: خبر المبتدأ الثاني مرفوع وعلامة رفعه ضمة ظاهرة في آخره وهو مضاف
	5. الجنة: مضاف إليه مجرور وعلامة جره كسرة ظاهرة في آخره.	5. الجنة: مضاف إليه مجرور وعلامة جره كسرة ظاهرة في آخره.
Percentage of Correct Answers	78.4%	74.5%

As shown in Table 4 ("Analysis of ChatGPT-4o on Grammatical Rules"), running ChatGPT-4o across different devices produced varied structural parsing outputs for the specified Arabic sentences. Nevertheless, ChatGPT-4o remains useful in facilitating learner language acquisition and proficiency development.

Using a sample of Arabic sentences, including “اللغة العربية هي لغة الجنة”, we obtained the percentage outcomes detailed in the table above. As shown, ChatGPT-4o achieved 78.4% accuracy on Device I and 74.5% on Device II, reflecting a minimal variance of 3.9%. This disparity is attributable to the underlying default system language configured on each device. Device I used Arabic as its primary interface language, improving information processing efficiency and the model's structural precision in presenting grammatical parsing. Conversely, Device II used an English interface, which influenced the conversational trajectory and led to variations in how I'rab analysis was presented.

4. Discussion

Analysis of Punctuation (*Harakat*) in Arabic

The findings indicate that GPT-4o received identical scores across the two accounts (Account I and Account II), scoring 4 on all four assessed indicators: accuracy of harakat pronunciation, clarity of phonetic separation, accuracy of *madd* rules, and consistency in recitation. Under the study's rating scheme, these scores indicate excellent performance. The absence of variation across accounts suggests that GPT-4o produced similarly rated outputs under the conditions examined. However, this finding should be interpreted as consistency between two evaluated accounts rather than evidence of universally high diacritical accuracy. Identical ratings do not necessarily mean the generated Arabic utterances were linguistically identical or error-free.

This distinction matters when interpreting the findings alongside previous empirical research on Arabic diacritization. For example, Qin et al. (2021) investigated Arabic diacritization using benchmark datasets. They showed that regularized decoding and adversarial training could improve performance even when some automatically generated linguistic information was inaccurate. Their findings illustrate that diacritization accuracy depends on how a system processes linguistic information and handles potentially noisy inputs (Favirotus Siyam et al. 2024; Qin et al. 2021).

Similarly, Abdel Karim and Abandah (2021) examined the effect of training-corpus size on automatic Arabic-text diacritization and reported improvements as the training data increased. Their best model achieved a diacritic error rate of 1.45% and a word error rate of 3.89% (Karim and Abandah 2021). Compared with these studies, the present research provides a different type of evidence: rather than measuring error rates against a benchmark corpus, it evaluates GPT-4o through expert-assessed performance categories. Although the present results are encouraging, the categorical scores cannot be directly compared with the error rates reported in these studies. The findings therefore complement existing research by examining account-level consistency, while leaving the actual frequency and distribution of diacritical errors undetermined.

The absence of observed differences between Account I and Account II may have several explanations. First, GPT-4o may have generated comparable outputs because both accounts received similar prompts, processed the same Arabic material, or operated under equivalent interaction conditions. Second, the model may have reproduced familiar linguistic patterns reliably for the specific words and sentence structures examined. Third, the rating instrument may not have been sufficiently sensitive to distinguish subtle differences between the outputs. For instance, two accounts could receive the same score despite differing in the placement of a single case-ending vowel, the articulation of a phoneme, or the duration of a long vowel. These possibilities are not mutually exclusive, and the present data do not establish which explanation is most plausible.

Research on broader Arabic-language capabilities supports this cautious interpretation. In GPTAraEval, Khondaker et al. (2023) evaluated ChatGPT and GPT-4 across 44 Arabic-language

understanding and generation tasks using more than 60 datasets. Their findings showed that, despite strong general-language capabilities, the evaluated models were consistently outperformed on several tasks by smaller models fine-tuned for Arabic; they also identified relative weaknesses in dialectal Arabic compared with Modern Standard Arabic (Khondaker et al., 2023). Although this study did not specifically establish GPT-4o's diacritization accuracy, it shows that model performance varies by task and language variety. Thus, the present study's identical scores should be understood within the tested conditions rather than treated as evidence that GPT-4o performs equally well across Classical Arabic, Modern Standard Arabic, dialectal varieties, and different levels of grammatical complexity.

Elgamal provides further relevant evidence in *Arabic Diacritics in the Wild*, which examines naturally occurring diacritics across several genres, including ChatGPT outputs. Their analysis identifies inconsistencies in the way ChatGPT handles diacritical forms, including variation in the ordering of certain diacritics and weaknesses in diacritical well-formedness. This research is especially pertinent because it demonstrates that the presence of diacritics in generated text does not automatically guarantee their linguistic correctness or consistent representation (Elgamal et al. 2024). However, their study addresses text-based diacritization rather than the precise speaking and recitation indicators assessed here. Accordingly, it should be used to contextualize the possibility of diacritical inconsistencies, not as direct evidence that the outputs in the present study were inaccurate. Taken together, the studies suggest that the performance of general-purpose language models should be evaluated at the level of the specific linguistic task, with particular attention to the difference between producing plausible Arabic and producing fully accurate Arabic.

Analysis of Arabic Pronunciation (Makhraj al-Huruf)

An Arabic linguistics expert evaluated and verified Chatbot GPT-4o's fluency and accuracy. These results demonstrate that the GPT-4o chatbot serves as an effective conversational partner for enhancing Arabic oral proficiency (Alhuwaydi 2024). The improvement in oral articulation quality facilitated by this chatbot aligns with empirical findings by Yuan (2024), which demonstrated that integrating interactive chatbots significantly enhances students' oral proficiency and willingness to communicate (WTC) (Yuan 2024).

Nevertheless, the GPT-4o chatbot's performance across Devices I and II was shaped by each device's default system language settings. The GPT-4o instance on Device I generated natural Arabic conversational output without requiring preliminary Indonesian prompting. This efficiency stems from Device I running with an Arabic system-language interface, which significantly streamlined evaluation of the chatbot's Arabic linguistic proficiency.

Conversely, the instance on Device II ran on an English interface. Consequently, this instance required preliminary prompts to steer the model toward generating Arabic conversational outputs. Nevertheless, GPT-4o on both Device I and Device II demonstrated robust Arabic conversational capabilities, achieving fluency that closely approximates native-speaker proficiency.

Pronunciation fluency and Arabic automatic speech recognition have advanced significantly. A recent study by Badawi et al. (2026) highlights that fine-tuning Automatic Speech Recognition (ASR) frameworks—such as OpenAI's Whisper—on expansive Arabic audio corpora substantially reduced the Word Error Rate (WER) from 35% to 7%, yielding considerably higher precision in assessing pronunciation accuracy and oral reading fluency (Badawi et al., 2026). Additionally, integrating ASR-based audio processing with Large Language Models (LLMs) has shown strong efficacy in capturing nuanced phonetic contexts and performing named entity recognition across complex acoustic environments (Boluk & Maratouq, 2026). These technological advancements bolster the role of AI chatbots as virtual tutors, a mechanism empirically shown to optimize

cognitive load management, stimulate intrinsic motivation, and enhance learner autonomy in language acquisition (Hong & Guo, 2025).

Analysis of ChatGPT-4o on Syntax (Ilmu Nahwu)

The findings in Table 3 indicate that ChatGPT-4o can provide accurate responses to basic nahw questions. At the same time, comparison with Al-Nahw Al-Wadih by Al-Jarimi suggests that its explanations can vary in structure without changing their core meaning. This flexibility matters for language learning because learners may benefit from explanations that present the same grammatical concept in different ways. The finding is consistent with Khondaker et al., whose comprehensive evaluation of ChatGPT across Arabic natural language processing tasks demonstrates its potential while also identifying limitations in Arabic-language performance (Khondaker et al. 2023). Similarly, Kwon et al. found that prompting strategies can improve large language models' performance in Arabic grammatical error correction, although instruction-tuned models do not consistently outperform systems specifically fine-tuned for the task (Kwon et al. 2023). These studies support interpreting ChatGPT-4o as a potentially useful resource for explaining basic grammatical concepts, rather than assuming that fluent and varied explanations guarantee accuracy across all areas of nahw.

From a pedagogical perspective, the ability to generate accessible explanations positions ChatGPT-4o as a potential supplementary tutor for Arabic grammar. Learners can use it to review concepts, request alternative explanations, and practice question-and-answer beyond the classroom. This interpretation aligns with Karataş et al., whose study of ChatGPT in foreign-language education identified potential benefits for grammar and vocabulary learning, as well as learner motivation and engagement (Karataş et al. 2024). Nevertheless, educational usefulness depends not only on the clarity of an explanation but also on its linguistic correctness and suitability for the learner's level. The present findings therefore suggest that ChatGPT-4o may facilitate access to basic nahw explanations. Still, they do not by themselves demonstrate improved learning outcomes or establish that the system can replace a qualified Arabic teacher.

However, the results should be interpreted within the scope of the tested questions. Correct responses to basic grammatical queries do not necessarily demonstrate reliable performance in more complex syntactic analysis, grammatical error correction, or context-dependent interpretation. In their Nahw benchmark, Mubarak et al. found substantial variation in large language models' Arabic grammar capabilities, with GPT-4o achieving an average score of 67% across the evaluated tasks (Mubarak, Hawasly, and Mohamed 2026). Likewise, Zbib et al. reported that language models can display strong surface-level proficiency while struggling with deeper grammatical and syntactic reasoning (Zbib et al., 2026). These findings highlight the importance of distinguishing the ability to produce plausible grammatical explanations from consistent mastery of Arabic grammatical rules. Accordingly, ChatGPT-4o is best regarded as a supplementary learning assistant whose responses should be checked against authoritative references such as Al-Nahw Al-Wadih and, where necessary, reviewed by an Arabic-language educator. Further research should test a broader range of grammatical topics and sentence structures, with repeated trials and independent expert assessment, to establish the consistency and reliability of its explanations.

Analysis: ChatGPT-4o on Grammatical Rules

The findings indicate that ChatGPT-4o achieved grammatical parsing accuracy rates of 78.4% on Device I and 74.5% on Device II, a 3.9-percentage-point difference. These results suggest that the model identified and explained several grammatical relationships in the Arabic sentence "اللغة العربية هي لغة الجنة", although it was not fully accurate. The difference between the two outputs is relatively small in descriptive terms, but it indicates that the generated *i'rab* analyses were not entirely

consistent across the tested conditions. This finding is consistent with Khondaker et al. (2023), whose comprehensive evaluation of ChatGPT across 44 Arabic natural language processing tasks found that performance varied across linguistic tasks and Arabic varieties, with Arabic-specialized models outperforming ChatGPT on several tasks (Khondaker et al., 2023).

More directly, benchmarking large language models on Arabic parsing found that GPT-4o achieved an F1 score of 0.83 in Arabic grammatical parsing under standardized prompting conditions, while also identifying persistent difficulties with complex grammatical constructions and diacritical ambiguity (Al-Sabahi et al. 2026). These studies reinforce the interpretation that ChatGPT-4o can support Arabic grammatical analysis, but its ability to generate plausible explanations should not be equated with consistently accurate syntactic interpretation. Accordingly, these findings provide preliminary evidence of useful grammatical performance while highlighting the need for expert verification, particularly when using the output for formal Arabic grammar instruction.

The difference between Device I and Device II may reflect differences in the language used during interaction, prompt formulation, or the conversational context established before the grammatical analysis. Nevertheless, the present results do not establish that the devices' default interface languages directly caused the 3.9-percentage-point difference. Because the comparison involved two distinct devices and accounts, the interface language may be confounded with other factors, including prompt wording, conversation history, account settings, and variations in the interaction sequence.

Evidence from Behzad et al., who asked LLMs about English grammaticality and then prompted them in a different language, shows that the language in which grammatical questions are presented can influence large language models' performance on metalinguistic tasks. However, their findings concern multilingual grammaticality judgments rather than *Arabic i'rab* analysis, so they support the plausibility of a prompt-language effect without proving the same mechanism in this study (Behzad, Zeldes, and Schneider 2024).

Similarly, Mubarak et al. found that large language models, including GPT-4o, still struggle with Arabic grammatical understanding across different evaluation tasks. Their findings suggest that differences in parsing accuracy may reflect the complexity of Arabic grammar and the model's interpretation of linguistic context, rather than interface language alone (Mubarak et al. 2026). Therefore, the higher score on Device I should be treated as an observed difference under the tested conditions, not as evidence that an Arabic-language interface inherently improves GPT-4o's grammatical reasoning.

Furthermore, the findings should be interpreted in light of the distinction between producing grammatically plausible Arabic explanations and demonstrating reliable grammatical competence. Elgamal explains that documented inconsistencies in ChatGPT-generated diacritics demonstrate that apparently appropriate Arabic output does not necessarily guarantee accurate diacritical representation (Elgamal et al. 2024).

Complementary evidence from Zbib reveals a gap between surface-level Arabic proficiency and deeper grammatical and syntactic understanding across evaluated language models. These studies contextualize the finding that ChatGPT-4o achieved accuracy rates below 80% on both devices, suggesting that its explanations may be educationally useful but still contain errors that require correction (Zbib et al., 2026). However, this evaluation is limited by its reliance on a single Arabic sentence, two device-account configurations, and the lack of reported repeated trials or a detailed item-level scoring procedure. The results therefore cannot establish whether the observed difference is statistically meaningful or generalizable to other Arabic sentence structures.

4. Conclusion

Testing ChatGPT-4o's performance across two distinct devices for spoken Arabic output shows significant potential alongside existing limitations. Based on the analysis of the two devices under study—ChatGPT-4o on Device I and Device II—both yielded satisfactory results in punctuation rendering and speech fluency, each achieving a score of 4 (Very Good). Furthermore, ChatGPT-4o's overall comprehension of Nahw (syntax) proved highly proficient; despite variations in sentence structure, the underlying core meaning remained intact. The observed divergence between the two ChatGPT-4o instances pertained to their ability to explain I'rab (syntactic parsing), with Device I outperforming Device II. Consequently, this study concludes that AI shows potential in Arabic conversational pronunciation, underscoring its viability as a practice tool for Arabic speaking skills. Nevertheless, human oversight remains essential to ensure precision and contextual understanding, particularly in Arabic language learning.

Despite these promising findings, this study has several notable limitations. Specifically, the evaluation was constrained by a single, simple Arabic sentence, a limited number of expert validators, and testing restricted to two specific devices. Given these constraints, further research is strongly recommended to address these gaps. Future research should examine AI's Arabic speech capabilities and grammatical mapping across a broader, more diverse corpus of sentences; incorporate a larger panel of linguistic experts; test performance across a wider array of operating systems and hardware devices; and explore multi-turn conversational interactions in complex linguistic contexts.

5. References

- Aidah Novianti Putri and Moh. Abdul Kholiq Hasan. 2022. 'Penerapan Kecerdasan Buatan sebagai Media Pembelajaran Bahasa Arab di Era Society 5.0'. *Tarling : Journal of Language Education* 7(1):69–80. doi:10.24090/tarling.v7i1.8501.
- Alfurqan, M. Hadli, & Wirdati Wirdati. (2024). 'Analisis Penggunaan Website Berbasis AI dalam Mengerjakan Tugas Akademik pada Mahasiswa Pendidikan Agama Islam Universitas Negeri Padang'. *ISLAMIKA* 6(4):1727–39. doi:10.36088/islamika.v6i4.5309.
- Alhuwaydi, A. M. (2024). 'Exploring the Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions – A Narrative Review for a Comprehensive Insight'. *Risk Management and Healthcare Policy* 17:1339–48.
- Allaithy, A., & Mai Zaki. (2025). 'Evaluation of AI-Generated Reading Comprehension Materials for Arabic Language Teaching'. *Computer Assisted Language Learning* 0(0):1–33. doi:10.1080/09588221.2025.2474037.
- Alruqi, T. N., & Salha M. Alzahrani. (2023). 'Evaluation of an Arabic Chatbot Based on Extractive Question-Answering Transfer Learning and Language Transformers'. *AI* 4(3):667–91. doi:10.3390/ai4030035.
- Al-Sabahi, K., Amer Fael Mohammed Balhaf, M. Moheb Zaghloul Al Shamy, Majjed Al-Qatf, Kang Yang, & Abdulrahman Al-Badwi. (2026). 'Benchmarking Large Language Models on Arabic Parsing'. *International Journal of Information Management Data Insights* 6(1):100404. doi:10.1016/j.jjime.2026.100404.
- Badawi, S. A., Maen Takruri, K. Salameh, Mohammad Al-Badawi, Nowar Alani, Isam ElBadawi, Aws Al-Qaisi, & Ghaleb Aldoboni. (2026). 'AI-Based Framework for Arabic Language Proficiency Assessment: A Deep Learning ASR Model with Enhanced Similarity Measures'. *Future Internet*, 18(5), 251. doi:10.3390/fi18050251.

- Behzad, S., Amir Zeldes, & Nathan Schneider. (2024). 'To Ask LLMs about English Grammaticality, Prompt Them in a Different Language'. Pp. 15622–34 in *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics.
- Boluk, Pinar, & Hana'a Maratouq. (2026). 'An AI-Based Security Architecture for Fraud Detection in Cloud Call Centers for Low-Resource Languages: Arabic as a Use Case'. *Electronics* 15(8):1718. doi:10.3390/electronics15081718.
- Chen, D.-L., Kirsi Aaltonen, H. Lampela, & Jaakko Kujala. (2024). 'The Design and Implementation of an Educational Chatbot with Personalized Adaptive Learning Features for Project Management Training'. *Technology, Knowledge and Learning*. doi:10.1007/s10758-024-09807-5.
- Creswell, John W., and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th edn. Los Angeles: Sage.
- Dokukina, I., & Julia Gumanova. (2020). 'The Rise of Chatbots – New Personal Assistants in Foreign Language Learning'. *Procedia Computer Science* 169:542–46. doi:https://doi.org/10.1016/j.procs.2020.02.212.
- Elgamal, S., Ossama Obeid, M. Kabbani, Go Inoue, & Nizar Habash. (2024). 'Arabic Diacritics in the Wild: Exploiting Opportunities for Improved Diacritization'. Pp. 14815–29 in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics.
- Favirotus Siyam, Fera, Rahmat Hidayat, Cecep Sobar Rochmat, Rosendah Dwi Maulaya, Annisa Avilya, & Muhammad Bahaudin Maulidi. (2024). 'Accuracy Analysis of Artificial Intelligence in Arabic Language Translation and Grammatical Rule Mapping'. *Jurnal AI Bayan: Jurnal Jurusan Pendidikan Bahasa Arab* 16(2):558. doi:10.24042/albayan.v16i2.24588.
- Fitri, W. A., & Muqita Hanifah Hasanah Dilia. (2024). 'Optimalisasi Teknologi AI dalam Meningkatkan Efektivitas Pembelajaran'. 5(10).
- Ghamrawi, N., T. Shal, & N. A. R. Ghamrawi. (2024). 'Exploring the Impact of AI on Teacher Leadership: Regressing or Expanding?' *Education and Information ...*
- Hong, X., & Li Guo. (2025). 'Effects of AI-Enhanced Multi-Display Language Teaching Systems on Learning Motivation, Cognitive Load Management, and Learner Autonomy'. *Education and Information Technologies* 30(12):17155–89. doi:10.1007/s10639-025-13472-1.
- Jeon, J., Seongyong Lee, & Hohsung Choe. (2023). 'Beyond ChatGPT: A Conceptual Framework and Systematic Review of Speech-Recognition Chatbots for Language Learning'. *Computers & Education*, 206:104898. doi:https://doi.org/10.1016/j.compedu.2023.104898.
- Karataş, F., Faramarz Yaşar Abedi, F. Özek Günyel, Derya Karadeniz, & Yasemin Kuzgun. (2024). 'Incorporating AI in Foreign Language Education: An Investigation into ChatGPT's Effect on Foreign Language Learners'. *Education and Information Technologies* 29(15):19343–66. doi:10.1007/s10639-024-12574-6.
- Karim, A. A., & Gheith Abandah. (2021). 'On the Training of Deep Neural Networks for Automatic Arabic-Text Diacritization'. *International Journal of Advanced Computer Science and Applications (IJACSA)* 12(8). doi:10.14569/IJACSA.2021.0120832.
- Khondaker, M. T. I., Abdul Waheed, E. Moatez Billah Nagoudi, & Muhammad Abdul-Mageed. (2023). 'GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP'. Pp. 220–47 in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics.
- Kwon, S., Gagan Bhatia, E. M. B. Nagoudi, & Muhammad Abdul-Mageed. (2023). 'Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction'. Pp. 101–19 in *Proceedings of ArabicNLP 2023*, edited by H. Sawaf, S. El-Beltagy, W. Zaghouani, W. Magdy, A. Abdelali,

- N. Tomeh, I. Abu Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, and R. Almatham. Singapore (Hybrid): Association for Computational Linguistics.
- Maulaya, R. D., Annisa Avilya, I. Syafa'atun Nur, Fasya Asy Shifaauzzahra, S. Fadhillah Y., & Fera Favirotus Siyam. (2024). 'Design Character Education Curriculum in Modern Pesantren Darussalam Gontor to Strengthen the Profile of Santri in Facing the VUCA Era'. *INTEGRATIA* 2(2):1-16.
- Mubarak, H., Majd Hawasly, & Abubakr Mohamed. (2026). 'Nahw: A Comprehensive Benchmark of Arabic Grammar Understanding, Error Detection, Correction, and Explanation'. Pp. 6310-28 in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by V. Demberg, K. Inui, and L. Marquez. Rabat, Morocco: Association for Computational Linguistics.
- Musthofa, A., Alfian, F. F. S., Alif Cahya Setiyadi, Y. Kusuma W., & Siti Aisyah. (2024). 'Efektivitas Media Pembelajaran Kotak Kartu Misterius (KoKaMi) untuk Meningkatkan Hasil Belajar Siswa pada Pelajaran Muthola'ah'. *Mantiqu Tayr: Journal of Arabic Language* 4(2):517-33. doi:10.25217/mantikutayr.v4i2.4671.
- Niyu, D. D., Azalia Gerungan, & Herman Purba. (2024). 'Penggunaan ChatGPT di Kalangan Mahasiswa dan Dosen Perguruan Tinggi di Indonesia'. *CoverAge: Journal of Strategic Communication* 14(2):130-45. doi:10.35814/coverage.v14i2.6058.
- Qin, H., Guimin Chen, Y. Tian, & Yan Song. (2021). 'Improving Arabic Diacritization with Regularized Decoding and Adversarial Training'. Pp. 534-42 in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, edited by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics.
- Regina Dwi Aulia, Shine Quinn Firdaus, Zaizafun Naura, & Nur Aini Rakhmawati. (2024). 'Analisis Pengaruh Penggunaan AI ChatGPT terhadap Minat Baca Mahasiswa Sistem Informasi IT'S'. *Jurnal Pendidikan, Bahasa dan Budaya* 3(3):01-11. doi:10.55606/jpbpb.v3i3.3196.
- Seyidov, R., & Ahmet Çitil. (2024). 'Arabic Speech Recognition and Artificial Intelligence Technologies'. *Journal of Ecohumanism* 3(4):937-42. doi:10.62754/joe.v3i4.3640.
- Suariqi Diantama. (2023). 'Pemanfaatan Artificial Intelligence (AI) dalam Dunia Pendidikan'. *DEWANTECH Jurnal Teknologi Pendidikan* 1(1):8-14. doi:10.61434/dewantech.v1i1.8.
- Wang, Y., Gai-Jing Wang, L.-J. Y., Jing Gao, Cong F., & Hai-Mei Ren. (2024). 'Qualitative Research on the Flipped Classroom Cognition of Undergraduate Nursing Students'. *BMC Medical Education* 24(1):1460. doi:10.1186/s12909-024-06426-7.
- Yuan, Y. (2024). 'An Empirical Study of the Efficacy of AI Chatbots for English as a Foreign Language Learning in Primary Education'. *Interactive Learning Environments* 32(10):6774-89. doi:10.1080/10494820.2023.2282112.
- Yulianti, Grace, Ngadi Permana, and Fitri Ayu Kusuma Wijayanti. (2023). 'Transformasi Pendidikan Indonesia: Memanfaatkan Potensi Kecerdasan Buatan (AI)'. 02(06).
- Zbib, M. Bilal, Hasan Abed Al Kader Hammoud, Ammar Mohanna, Nadine Rizk, Fatima Karnib, Sina Moukaled, & Bernard Ghanem. (2026). 'AraLingBench: A Human-Annotated Benchmark for Evaluating Arabic Linguistic Capabilities of Large Language Models'. Pp. 385-93 in *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, edited by M. El-Haj, P. Rayson, M. Jarrar, I. Ezeani, S. Ezzini, S. Ahmadi, A. Haddad Haddad, C. Amol, A. Abdelali, and S. Abudalfa. Rabat, Morocco: Association for Computational Linguistics.
- Zhang, Z., & Xiaomeng Huang. (2024). 'The Impact of Chatbots Based on Large Language Models on Second Language Vocabulary Acquisition'. *Heliyon* 10(3):e25370. doi:10.1016/j.heliyon.2024.e25370.

الأول. الأولى. بوروت: مكتبة منيمه. Vol. علم النحو والصرف. عتيق, الدكتور عبد العزيز