Acceptability and Effectiveness Analysis of Large Language Model-Based Artificial Intelligence Chatbot Among Arabic Learners

Nely Rahmawati Zaimah[1]*, Eko Budi Hartanto[2], Fatchiatu Zahro[3]
[1]Arabic Education Study Program Sekolah Tinggi Agama Islam Al-Anwar Rembang, Indonesia.
[2]Arabic Education Study Program Institut Agama Islam Negeri Kediri, Indonesia.
[3]Arabic Education Study Program Institut Ummul Qurro Al-Islami Bogor, Indonesia.
Correspondence Address: neyrzah@staialanwar.ac.id

Received: 10-10-2023	Revised: 22-11-2023	Accepted: 05-12-2023

Abstract

This research stems from the broad use of AI based on Large Language Models (LLMs), which many academics find relevant and effective in higher education Arabic language learning. The goal is to confirm these views.This research is a mixed reseach that employs a both of qualitative and quantitative methodologies. The qualitative segment involves observations and literature reviews. Observations involved reviewing how participants used chatbots and carefully checking the accuracy and consistency of platform responses. The quantitative facet utilizes a paired experimental design, encompassing both classical and Bayesian Paired Sample t-Tests analysis. The research encompasses 45 individuals with a proficient understanding of Modern Standard Arabic and no hindrances in comprehending the material. These individuals are enrolled as students at Islamic College (STAI) Al-Anwar Rembang, Indonesia. The results show increased motivation and ease of use with the chatbot in Arabic language learning. However, concerns about the consistency of chatbot content have arisen, affecting participants' confidence in response accuracy of AI. This prompts an evaluation of effectiveness through classical and Bayesian tests, which fail to demonstrate statistically significant variances, even in the adaptive Bayesian probability analysis. These outcomes deviate from previous research on relevance and effectiveness and corroborate preceding studies on academic apprehensions and accuracy enhancements. The researchers advocate for further investigations, especially concerning the accuracy analysis of AI chatbots in Arabic pedagogical contexts.

Keywords: Artificial Intelligence Chatbot, Large Language Model, Non-Native Arabic Learners

ملخص

ينبع هذا البحث من الاستخدام الواسع للذكاء الاصطناعي القائم على نماذج اللغة الكبيرة، الذي يجده العديد من الأكاديميين ذا ذو صلة وفعالية في تعلم اللغة العربية في التعليم العالي. الهدف هو تأكيد هذه الآراء. هذا البحث هو بحث مختلط يعتمد على منهجيات كمية ونوعية على حد سواء. يتضمن الجانب النوعي المراقبات واستعراضات أدبية. تمت المراقبات عن طريق مراجعة كيفية استخدام المشاركين للروبوتات الدردشة وفحص دقيق لدقة واتساق استجابات المنصات. يستخدم الجانب الكمي تصميمًا تجريبيًا متزاوجًا ، يغطي كل من التحليل الكلاسيكي واختبارات العينات المتزاوجة باستخدام النماذج البيانية. يشمل البحث ٤٥ فردًا لديهم فهم جيد للعربية الفصحى ولا يواجهون عقبات في فهم المواد. هؤلاء الأفراد مسجلون كطلاب في كلية الإسلامية الأنور في رمبانغ، إندونيسيا. تظهر النتائج زيادة في الحماس وسهولة الاستخدام مع ميزة الدردشة في تعلم اللغة العربية. ومع ذلك، ظهرت مخاوف بشأن اتساق محتوى الدردشة، مما أثر على ثقة المشاركين في دقة ردود الدردشة. وهذا يدفع إلى تقييم الفعالية من خلال اختبارات العينات المتزاوجة الكلاسيكية والبيانية، التي لا تظهر اختلافات ذات دلالة إحصائية، حتى في تحليل الاحتمال البياني التكيفي. تتباين هذه النتائج عن الأبحاث السابقة بشأن الصلة والفعالية وتؤكد الدراسات السابقة حول مخاوف الأكاديميين

DOI: 10.25217/mantiqutayr.v4i1.3951

وتحسين الدقة. يشجع الباحثون على إجراء مزيد من البحوث، خاصة فيما يتعلق بتحليل الدقة للروبوتات الذكية في سياق التعليم العربي.

الكلمات المفتاحية: روبوت ذكاء صناعي، متعلمون عربية غير أصليين، نموذج لغوي كبير

Introduction

The utilization of Artificial Intelligence chatbots has seen rapid growth in several years and has become a new trend in virtual communication. Its usage in any fields has become an interesting research topic and is still under development in terms of its positive or negative aspects.[1] The emergence of Large Language Models (LLMs) such as Bing chatbot and ChatGPT has opened up new opportunities for creating more adaptive and interactive learning experiences.[2]In addition to the numerous studies on its effectiveness and progress in various aspects, there are also many studies that encourage new considerations and caution in the use of AI, for example in the fields of medicine, religion, and linguistics.[3]

This research arises as a response to the growing use of Artificial Intelligence (AI) and the increasing recognition of its utility as an aid or resource in various academic fields. The relevance and effectiveness of AI have drawn the attention of many scholars and experts, who see it as a significant influence in the educational landscape. The main motivation behind this research is to validate these views in the context of Arabic language learning in higher education. It's about understanding AI's position and relevance in Arabic language learning and to what extent it influences learners. As AI continues to advance, especially in the field of language learning, this research becomes a necessary step.

However, the context of the Arabic language differs from other educational landscapes. With all its intricacies and rich linguistic heritage, it serves as the primary backdrop for this exploration of the relevance of AI chatbots in Arabic language studies. The complexities and variations within the Arabic language make it a challenging yet highly valuable subject to study. Understanding how AI

[1] Aqil M Azmi, Abdulaziz O Al-Qabbany, and Amir Hussain, *"Computational and Natural Language Processing Based Studies of Hadith Literature: A Survey,"* Artificial Intelligence Review 52 (2019): 1384-85. DOI: https://doi.org/10.1007/s10462-019-09692-w; Luigi De Angelis et al., *"Chat GPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health,"* Frontiers in Public Health 11 (April 25, 2023): 1166120. DOI: https://doi.org/10.3389/fpubh.2023.1166120; Ben Shneiderman, *"Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy,"* International Journal of Human–Computer Interaction 36, no. 6 (April 2, 2020): 499. DOI: https://doi.org/10.1080/10447318.2020.1741118

[2] Fitrah Rumaisa et al., *"Penerapan Natural Language Processing (NLP) di Bidang Pendidikan,"* Jurnal Inovasi Masyarakat 1, no. 3 (2021): 232–35. DOI: https://doi.org/10.33197/jim.vol1.iss3.2021.799; Thomas KF Chiu et al., *"Teacher Support and Student Motivation to Learn with Artificial Intelligence (AI) Based Chatbot,"* Interactive Learning Environments, 2023, 12-13. DOI: https://doi.org/10.1080/10494820.2023.2172044

[3] Joshua James Hatherley, *"Limits of Trust in Medical AI,"* Journal of Medical Ethics 46, no. 7 (2020): 479. DOI: https://dx.doi.org/10.1136/medethics-2019-105935; Douglas Johnson et al., *"Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model,"* Research Square 2023. DOI: https://doi.org/10.21203/rs.3.rs-2566942/v1

can be utilized to facilitate Arabic language learning is not just an academic pursuit but also a practical need for educators and learners.[4]

In essence, this research aims to bridge the gap between the theoretical potential of AI in language education and its practical application in the unique context of Arabic language learning. It seeks to provide insights into AI's position and relevance in the field of Arabic language learning in higher institutions, shedding light on the extent to which it impacts learners. As AI continues to evolve, this research serves as a crucial step in understanding its role in Arabic language education in higher institutionsjust in line withRitonga et al., in implementational systematization.[5]

The main goal of this research is to assess the level of acceptability of AI chatbots that use Large Language Models (LLMs) among Arabic language learners as a second language (L2) in Indonesia, within the context of Arabic language learning. This study also aims to identify the factors that influence the acceptance of AI chatbots based on LLMs as a learning tool for non-native speakers of Arabic and measure its effectiveness. These factors include motivation-enthusiasm[6], learners' beliefs[7], and the validity of using AI in education, making them instruments of measurement and analysis themes.

Furthermore, this research intends to conduct a comparative analysis of its findings with recent research in the field, which has shown its effectiveness[8], although there are also concerns[9], especially in the medical, psychological and educational landscapes. Thus, this research aims to determine whether its findings align with or contradict recent research, contributing to the current discussion about AI chatbots and their role in Arabic language learning. This comparative analysis serves as a means to highlight any updates or novelty in the findings and to establish the research's position within the current academic landscape.

Artificial Intelligence has brought many new opportunities to the way society functions and in the dynamics of education.[10] The extensive use of AI has separated humanity into two groups: the

[4] Munir Munir, *"Pendekatan Struktural dalam Pelajaran Bahasa Arab,"* Shaut al Arabiyyah 6, no. 1 (August 24, 2018): 13, DOI: https://doi.org/10.24252/saa.v6i1.5644; Islamiyah Sulaeman, Syuhadak Syuhadak, and Insyirah Sulaeman, *"ChatGPT as a New Frontier in Arabic Education Technology,"* Al-Arabi: Jurnal Bahasa Arab dan Pengajarannya= Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: https://dx.doi.org/10.17977/um056v7i1p83-105; Saiful Anwar, Guntur Cahaya Kesuma, and Koderi, *"Development of Al-Qawaid an-Nahwiyah Learning Module Based on Qiyasiyah Method for Arabic Language Education Department Students | Pengembangan Modul Pembelajaran al-Qawaid an-Nahwiyah Berbasis Metode Qiyasiyah untuk Mahasiswa Jurusan Pendidikan Bahasa Arab,"* Mantiqu Tayr: Journal of Arabic Language 3, no. 1 (January 2, 2023): 11–24. DOI: https://doi.org/10.25217/mantiqutayr.v3i1.2830

[5] Mahyudin Ritonga et al., *"Analysis of Arabic Language Learning at Higher Education Institutions with Multi-Religion Students,"* Universal Journal of Educational Research 8 (September 1, 2020): 4333–39. DOI: https://doi.org/10.13189/ujer.2020.080960

[6] Mita Rosyda Attaqiana, Saptorini Saptorini, and Achmad Binadja, *"Pengembangan Media Permainan Truth and Dare Bervisi Sets Guna Memotivasi Belajar Siswa,"* Jurnal Inovasi Pendidikan Kimia 10, no. 2 (2016) : 1798 – 1806. DOI: https://doi.org/10.15294/jipk.v10i2.9533

[7] Achmad Hidayatullah and Csaba Csíkos, *"The Role of Students' Beliefs, Parents' Educational Level, and The Mediating Role of Attitude and Motivation in Students' Mathematics Achievement,"* The Asia-Pacific Education Researcher, March 30, 2023. DOI: https://doi.org/10.1007/s40299-023-00724-2.

[8] Ahlam Fuad and Maha Al-Yahya, *"Recent Developments in Arabic Conversational AI: A Literature Review,"* IEEE Access 10 (2022): 23842. DOI: https://doi.org/10.1109/ACCESS.2022.3155521

[9] Mike Perkins, *"Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond,"* Journal of University Teaching and Learning Practice, British University, Vietnam 20, no. 2 (February 22, 2023). DOI: https://doi.org/10.53761/1.20.02.07; Miles Brundage et al., *"Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,"* arXiv Preprint arXiv:2004.07213, 2020. DOI: https://doi.org/10.48550/arXiv.2004.07213

[10] Amr M Mohamed, *"Exploring the Potential of an AI-Based Chatbot (ChatGPT) in Enhancing English as a Foreign Language (EFL) Teaching: Perceptions of EFL Faculty Members,"* Education and Information Technologies, 2023: 11. DOI: https://dx.doi.org/10.1007/s10639-023-11917-z

progressive ones who are enthusiastic about using and developing AI in their respective fields, and on the other side, the concerned ones who request careful consideration before further entanglement.[11] In the realm of academic research, there has been a conflict between these two camps. Many researchers have even advocated for institutional or ethical policies and further scrutinized its negative aspects.The Arabic language, which is directly linked to social, cultural, and religious studies, is highly sensitive to this issue. Its application in the learning of the Arabic language and culture is also highlighted.[12]

Positively, the research conducted by Abdulkader and Al-Irhayimin[13] emphasize that Arabic, with its unique characteristics and various variations, including Classical Arabic (CAL), Modern Standard Arabic (MSA), and Arabic dialects (DA), used practically differently in various contexts, can be integrated into various chatbots as intelligent technology using Artificial Intelligence (AI) to communicate with humans (in this case, language learners) in their natural language. The primary function of chatbots is to understand user requests and provide the most appropriate responses using Natural Language Processing (NLP) techniques.[14] These findings reflect the importance of expanding the scope of technology and measuring its effectiveness for pedagogical and language education purposes.

Additionally, Fuad and Yahya's[15] paper discussing the introduction of AI in the context of conversational Arabic language learning illustrates the significant potential of using AI-based chatbots for language learning but also acknowledges the need to understand how learners accept this technology for learning Arabic.In line with that, the research conducted by Rumaisha titled "Application of Natural Language Processing (NLP) in Education," and the research by Chiu et al. on "Teacher Support and Student Motivation to Learn with an Artificial Intelligence (AI) Based Chatbot"[16] highlight the effectiveness and relevance of AI chatbots in the field of education. Furthermore, Shao et al.,[17] emphasize that teaching Arabic as an L2 in an online or technology-based environment will have a significant impact on speaking skills in Arabic and has become a new

[11] Joshua James Hatherley, *"Limits of Trust in Medical AI,"* Journal of Medical Ethics 46, no. 7 (2020): 478–81. DOI: https://dx.doi.org/10.1136/medethics-2019-105935; Douglas Johnson et al., *"Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model,"* Research Square 2023. https://doi.org/10.21203/rs.3.rs-2566942/v1

[12] Ismail Suardi Wekke and Maimun Aqsha Lubis, *"A Multicultural Approach in Arabic Language Teaching: Creating Equality at Indonesian Pesantren Classroom Life,"* Sosiohumanika 1, no. 2 (2008): 296-310. DOI: https://doi.org/10.2121/sosiohumanika.v1i2.337; Islamiyah Sulaeman, Syuhadak Syuhadak, and Insyirah Sulaeman, *"ChatGPT as a New Frontier in Arabic Education Technology,"* Al-Arabi: Jurnal Bahasa Arab dan Pengajarannya= Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: http://dx.doi.org/10.17977/um056v7i1p83-105

[13] Zena Abdulkader and Yousra Al-Irhayim, *"A Review of Arabic Intelligent Chatbots: Developments and Challenges,"* Al-Rafidain Engineering Journal (AREJ) 27, no. 2 (September 1, 2022): 178–89. DOI: https://doi.org/10.33899/rengj.2022.132550.1148

[14] Fitrah Rumaisa et al., *"Penerapan Natural Language Processing (NLP) di Bidang Pendidikan,"* Jurnal Inovasi Masyarakat 1, no. 3 (2021): 233. DOI: https://doi.org/10.33197/jim.vol1.iss3.2021.799

[15] Ahlam Fuad and Maha Al-Yahya, *"Recent Developments in Arabic Conversational AI"* IEEE Access Volume: 10 (2022): 23842 - 23859. DOI: https://doi.org/10.1109/ACCESS.2022.3155521; Erham Budi Wiranto and Sri Suwartini, *"Artificial Intelligence and Trustworthy Principles in Global Islamic Education,"* Ushuluddin International Conference (USICON) 6 (2022): 80. https://conference.uin-suka.ac.id/index.php/USICON/article/view/1252

[16] Thomas KF Chiu et al., *"Teacher Support and Student Motivation to Learn with Artificial Intelligence (AI) Based Chatbot,"* Interactive Learning Environments 2023: 8. DOI: https://doi.org/10.1080/10494820.2023.2172044

[17] Sicong Shao et al., *"AI-Based Arabic Language and Speech Tutor,"* in *2022 IEEE/ACS* 19th International Conference on Computer Systems and Applications (AICCSA) (2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates: IEEE, 2022), 1–8. DOI: https://doi.org/10.1109/AICCSA56895.2022.10017924

phenomenon. However, the above-mentioned research does not delve further into examining the experts' acceptance regarding language norms and contextual precision that supported by empirical studies. This aspect of acceptance will undoubtedly correlate with the study's outcomes.

Therefore, an analysis of the acceptance of LLMs-based chatbots in Arabic language learning will provide valuable insights into how this technology can be effectively integrated, especially in Arabic language learning. With various advantages, there are no longer any limitations in developing AI-based learning, even in challenging situations like the COVID-19 pandemic, as well as Arabic Language in open and distributed learning context.[18]

On the other hand, De Angelis et al.in the research tittle "ChatGPT and the Rise of Large Language Models: the New AI-Driven Infodemic Threat in Public Pealth" and also Perkins[19]in "Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond" emphasize the importance of rapid policy development to address potential threats and ethical issues. One of the main challenges is the difficulty in accurately detecting text generated by artificial intelligence. LLMs can quickly generate a large amount of text, which can be used to spread misinformation or misleading information on an unprecedented scale.

Previously, in the medical field, both Hatherley[20] and Gilbert et al.[21]indicate thatgenerative AI has been questioned for its accuracy and reliability. However, technology based on LLMs is primarily a language communication development model, which is not predominantly designed to consider scientific precision and the empirical spirit of science. So, how can its use be reliable in constructing educational frameworks? Further research is needed in various measurement variables, including acceptance and effectiveness aspects, as in this study.There is a need for in-depth analysis of its usage, especially in determining its relevance and linguistic validity. Therefore, further studies and research are needed to develop tools and formulate policies to address these gaps.

The authors attempt to explore the acceptability factors and analyze them in the context of Arabic language learning, where there are strict constructions regarding proper Arabic language techniques and practices, as reviewed by several previous researchers, whether curriculum-based or using various approaches. There are several acceptability parameters to be analyzed, including Usability, Effectiveness, and Validity parameters. The Usability parameter measures how easy users find it to use AI chatbots. This includes reviews measuring the ease of interaction with leading AI chatbots, especially those with interfaces and features in Indonesian and Arabic for communication. Further analysis will assess whether the chatbot interface is intuitive and how quickly learners can master its use. The Effectiveness parameter is used to measure whether the use of AI chatbots based on LLMs actually improves learners' Arabic language proficiency. The Authors measures their progress before and after using chatbots to see how effective they are in supporting learning. The Validity evaluation parameter[22] can be used to analyze the standardization of these chatbots in

[18] Nita Amelia, (second) Noval Mulya Dava S., and (third) Muhammad Minan Chusni, "Pemanfaatan Artificial Intelligence Dalam PembelajaranDi Masa Pandemi| Prosiding Amal Insani Foundation," September 6, 2023, https://prosiding.amalinsani.org/index.php/semnas/article/view/10

[19] De Angelis et al., "ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health," *Frontiers in Public Health* 11 (April 25, 2023): 1166120. DOI: https://doi.org/10.3389/fpubh.2023.1166120

[20] Joshua James Hatherley, "Limits of Trust in Medical AI," *Journal of Medical Ethics* 46, no. 7 (2020): 479-480. DOI: https://dx.doi.org/10.1136/medethics-2019-105935

[21] Stephen Gilbert et al., "Large Language Model AI Chatbots Require Approval as Medical Devices," *Nature Medicine* 29, no. 10 (October 2023): 2396–98. DOI: https://doi.org/10.1038/s41591-023-02412-6

[22] Cecilia Ka Yuk Chan, *Assessment for Experiential Learning,* (1st ed.). Routledge London 29 September (2022): 378. DOI: https://doi.org/10.4324/9781003018391

linguistic terms, such as their recognition of learners' language style and adaptation, and whether there is a risk of diglossic spelling and grammar violations.[23] By formulating these parameters, researchers will have a strong framework for analyzing the acceptability of AI chatbots based on LLMs in Arabic language learning.

Lastly, the research findings will provide valuable input into the scientific knowledge related to the capabilities of AI, its usage, accuracy, level of acceptance, and effectiveness in the field of education, especially its applicability in Arabic language learning. Of course, time limitations and research subjectivity are also potential gaps. However, the authors hope for new research in the scope of learning with this technology to further expand the horizons of knowledge in this field.

Method

The research employs a two-fold methodology, mixed both qualitative and quantitative approaches, to gauge the acceptability of AI chatbots based on LLMs in the context of Arabic language learning as an L2. The qualitative approach involves assessing the participants' initial performance by observing their average grades before commencing the Arabic language course. This data will be leveraged to identify patterns and trends in how participants interact with AI chatbots. Additionally, a brief questionnaire will be administered to delve into the motivations driving individuals to learn Arabic with the assistance of AI chatbots.

On the other hand, the quantitative approach adopts an experimental design, utilizing a paired sample t-test to measure the initial effectiveness after distribution testing. Quantitative data will be collected through pre-tests and post-tests containing relevant Arabic language tasks.[24] The participants in the experiment will consist of 45 students, encompassing both males and females, from STAI Al-Anwar Sarang Rembang. These students are enrolled in Arabic language classes (IQT) and possess a sufficient level of proficiency in comprehending Modern Standard Arabic (MSA) with an average background of traditional pesantren students who have been studying Standard Arabic for a long time.

The research design revolves around observing the acceptance levels in Arabic language learning through AI chatbots. Following this, an experimental phase is initiated employing paired sample t-tests, aimed at evaluating the initial effectiveness post-distribution testing. The data derived from these tests will undergo analysis, which can be executed using either the conventional paired sample t-test or Bayesian inference models. This analysis aims to compare the disparities between pre-test and post-test scores in Arabic language tasks, complete with their respective probability values. This comprehensive evaluation facilitates the measurement of the AI chatbots' effectiveness in enhancing the Arabic language skills of the participants. For this analysis, JASP 0.18.0.0 software is employed due to its proficiency in elucidating Bayesian factors and their associated posterior distribution.[25]

[23] Risa Rahmah, Azizatul Muzdalifah, & Mu'alim Wijaya. *"Penggunaan Thariqah Mubasyarah sebagai Pembelajaran Bahasa Arab yang Efektif."* Al Maghazi : Arabic Language in Higher Education, 1.1 (2023): 23-29. DOI: https://doi.org/10.51278/al.v1i1.706

[24] Mustaqim Mustaqim, *"Metode Penelitian Gabungan Kuantitatif Kualitatif/Mixed Methods Suatu Pendekatan Alternatif,"* Intelegensia: Jurnal Pendidikan Islam 4, no. 1 (2016): 1-9. DOI: https://doi.org/10.34001/intelegensia.v6i1.1351

[25] Wei Liang and Hongsheng Dai, *"Bayesian Inference,"* in Quantum Chemistry in the Age of Machine Learning (Elsevier, 2023): 240-42. DOI: https://doi.org/10.1016/B978-0-323-90049-2.00005-6

Result and Discussion

Moving Beyond the Acceptance Observation

The process of collecting qualitative data began by gathering academic records from previous years found on the official website of STAI Al-Anwar Sarang Rembang. Out of the 45 students enrolled in the Department of Quranic Studies (IQT), there was significant variation in their academic performance. The researcher also conducted an initial assessment of their competence in various aspects of the Arabic language, including conversation, grammar, vocabulary, and presentation skills. Among the 45 students, 27 of them scored above average and exceeded the threshold set by the researcher beforehand (5.50). Their average score was 6.81. There were 12 students within the score range close to the threshold (between 5.51 to 6.80), 1 student had a score equal to the threshold, while 5 students scored below the threshold. Therefore, the assessment of these skills qualified them for participation in the research.

Five questions were posed to all participants during the observation phase and questionnaire filling, resulting in enthusiastic responses. Only minor notes were added as additional information, with an emphasis on guidance provided by the researcher and instructors to optimize AI chatbot usage using Bing chat and ChatGPT. The observations and questionnaires provided valuable insights into the participants' initial conditions and their readiness to interact with AI chatbots in Arabic language learning. Academic records and competence assessments played a crucial role in identifying variations in the language abilities of the students. The questionnaires administered to all participants yielded positive and enthusiastic responses. Participants showed a desire and eagerness to interact with AI chatbots in their language learning journey. The minor notes added to the responses emphasized the valuable guidance provided by the researcher and instructors, with a focus on using Bing chat and ChatGPT to maximize AI chatbot capabilities. In the initial stage of this research, the researcher successfully measured the participants' initial language abilities, paving the way for the next phase of the research, which will involve a quantitative assessment of the effectiveness of AI chatbots in improving their Arabic language skills. This information is crucial for designing and implementing effective language learning experiences with AI chatbots as valuable learning tools.

The researcher conducted a brief training session on the use of AI chatbots in Arabic language learning, aiming to provide students with a strong understanding of how to integrate AI chatbot technology into the Arabic language learning process, including its intricacies. This training will include various steps that will enable participants to effectively utilize AI chatbots in enhancing their Arabic language skills. The duration of this brief training session is one hour. The duration may expand during the utilization and assessment sessions, allowing participants to gain a better understanding of the concept of AI chatbots in Arabic language learning. This training incorporates various learning approaches and categories, including presentations, direct demonstrations, practice sessions, and discussions in conversation, writing, and interpretation panels. Each participant will have access to their own personal AI chatbot account for direct practice. The brief training session is expected to equip participants with the necessary skills and understanding to maximize the use of AI chatbots in their Arabic language learning. The integration of AI chatbots into Arabic language learning sessions will be conducted over two subsequent meetings, including practical exercises. Afterward, the researcher will administer utilization tasks or exams consisting of three assessment categories: conversational fluency, writing skills, and interpretation.

In the integration phase of Arabic language learning, the first step will begin with a session of easy conversation with Arabic Standard (MSA) and Classical chatbots, freely discussing anything related to *Mubtada'* (the subject or the noun that begins a sentence) and *Khobar* (the predicate or the part of the sentence that provides information about the subject) materials and scrutinizing every response provided by ChatGPT and Bing Chat. All of this will be conducted in Arabic. Out of the 45 students, they will be organized into 5 groups, each containing 8 and more participants, with 5 question models focusing on the same material, which will be further developed based on each chatbot's responses. The responses provided by the chatbot vary. The first group's theme is the definition of *Mubtada'* and *Khobar*, and each participant follows the chatbot's responses while the researcher allows them to discuss with their group members. The answers accompanying each participant are different. The second group focuses on the types of *Mubtada'* and *Khobar* along with examples. All participants receive nearly the same answers, albeit in different phrasing, and then collectively discuss them. The third group creates questions related to examples of the development of *Mubtada-Khobar* from *ism dhomir* and *ism dhohir*, and all participants receive different answers. The fourth group asks the chatbot about *Mubtada'* and *Khobar*, focusing on *Mufrod* and *Ghoyru Mufrod*. In the development of their responses, all participants receive identical answers but with varying phrasing. The last group, with the highest average score, focuses on fact-based questions derived from *Mubtada'* and *Khobar*, regardless of type, and takes examples of *ghoryb* or unique instances rarely found but sometimes encountered in classical literature. Almost all participants receive different answers. Finally, the researcher provides a key question for the interactive comprehension level, material absorption, and acceptability (in the form of criticism and feedback), which will then be assessed numerically.

In the subsequent integrative meetings, a more complex scheme is introduced with a final assessment for the same categories. The same groups are assigned different themes. The first group discusses *An-na'tal-haqîqy* (the real or essential predicate) and *An-na't As-sababy* (the attributive predicate) and follows the chatbot's responses. The second group focuses on *At-tawkid*. The third group deals with Al-Hal and the dynamics of their responses. The fourth group explores *At-tamyiz* with its models and rules. The fifth group delves into differentiating factors of *Adawātul-Istifhām* and anything that enhances their understanding in the chatbot. The researcher also evaluates them in the same categories as before.

The third session involves translating a paraphrase from Arabic using the chatbot and then reading it aloud. This phase assesses participants' understanding and critical thinking. Differences in translation between the chatbot and other translation platforms, both literally and contextually, are examined. At this stage, participants encounter difficulties due to variations in the chatbot's responses. Group 1, which tends to provide more literal translations, may have difficulties understanding and recognizing figurative language (*majāz*) in Arabic texts. The discussion of *majāz* rules provides a better understanding of how *majāz* can be used in Arabic to convey deeper meanings or for rhetorical effects. Group 2, which tends to provide more contextual translations, is more open to discussing *majāz*. They have started to recognize the presence of *majāz* in Arabic texts and try to understand them. This discussion helps them better understand how *majāz* can be used to express complex and profound meanings in Arabic. Groups 3 and 4, with variations in their translation approaches, also benefit from the discussion of *majāz* rules. They begin to understand that the use of *majāz* can make Arabic texts more beautiful and expressive. Group 5, which also strives to reflect nuances in their translations, finds that the discussion of *majāz* rules helps them appreciate the beauty of the Arabic language and its ability to convey meaning creatively. They begin to recognize various *majāz* and how they can be used to express deep and complex meanings. The discussion of *majāz* rules provides deeper insights to all participant groups about how Arabic is used rhetorically and artistically in classical literature. This helps them better understand nuances in Arabic texts and recognize the use of *majāz* in the language. With a better understanding of *majāz* rules, participants can become more skilled readers and translators in Arabic. The researcher also awards extra points because participants have a non-material mastery of linguistic understanding in a broader context.

Based on the observations and monitoring of using a chatbot in Arabic language learning, we have uncovered several crucial findings regarding usability, user comfort, and the validity of utilizing this chatbot in the context of Arabic language learning. In this discussion, we will delve deeper into these discoveries.

First and foremost, concerning the chatbot's usability in Arabic language learning, our findings from observations and monitoring indicate that the chatbot can serve as a relatively user-friendly tool for learners. This is attributed to the chatbot's intuitive interface and its well-designed, user-friendly features. Learners, especially those who are tech-savvy, tend to feel at ease while interacting with the chatbot to enhance their Arabic language skills. However, it's important to note a few critical considerations. Firstly, the chatbot's usability may vary depending on the learners' initial proficiency levels. Learners with a basic understanding of Arabic may find it more accessible to adapt to the chatbot compared to complete beginners. Therefore, there should be adjustments in the instructional design to cater to learners with varying proficiency levels. Additionally, continuous improvement and development of the chatbot are vital to make it more responsive to users' requirements. In the realm of language learning, the chatbot should have the ability to detect learners' errors, offer constructive feedback, and customize learning materials based on individual progress. These measures will enhance usability and effectiveness.

Regarding comfortability in using the chatbot as a tool for Arabic language learning, most participants feel comfortable interacting with the chatbot during the learning process. They perceive the chatbot as a helpful learning partner and not intimidating. However, there are some aspects of comfort that need attention. Some participants may experience slight anxiety or discomfort when interacting with technology, especially if they are not tech-savvy users. Therefore, additional support in the form of user guides or resources that can help participants feel more comfortable is needed. Additionally, the chatbot should create an inclusive and friendly learning environment. It should not make participants feel pressured or afraid to make mistakes. Instead, the chatbot should encourage language experimentation and provide constructive feedback.

Lastly, in terms of the validity of using the chatbot in Arabic language learning. Validity is a measure of how well the chatbot can assess what should be assessed, which is the participants' Arabic language proficiency. Findings indicate that the chatbot has the potential to assess participants' Arabic language proficiency effectively. However, the validity of the chatbot can be influenced by several factors. First, it is important to ensure that the chatbot assesses various aspects of Arabic language, including comprehension, speaking, reading, and writing. Additionally, the chatbot should have evaluation items that comprehensively measure language proficiency in many terms. Validity can also be enhanced by ensuring that the chatbot provides accurate and informative feedback to participants. This feedback should help participants understand their errors and provide guidance for improvement. Interestingly, different responses emerged from each task presented to the participants. These varying responses also corresponded to the columns for question corrections (1, 2, etc.). This raised doubts about the accuracy of the chatbot regarding the participants' mastery of the subject matter and somewhat diminished their motivation. Some consider it just for fun, while others are afraid to develop it further.

Effectiveness Hindered by Concerns

The results of the assessment for each participant within the groups can be described as table and plots follows:

Table 1. Descriptives ofthe Assessment Result

	N	Mean	SD	SE	Coefficient of variation
Rata-rata Sbl	45	6.457	1.164	0.174	0.180
Rata-rata	45	6.562	1.000	0.149	0.152

Note: The descriptive result by JASP 0.18.0.0.

Based on table 1 it is found that the provides information about the comparison of average values before and after an experiment or intervention involving 45 participants. Before the intervention, the average score was 6.457 with a standard deviation of 1.164, while after the intervention, the average score increased to 6.562 with a standard deviation of 1.000.This description table indicates that the intervention or experiment may have had a positive impact, as the participants' average scores increased. The lower standard deviation in the scores after the intervention suggests that the data tends to be more concentrated around the mean, which can be interpreted as an increase in consistency in the results.There is a change in the Standard Error (SE), indicating the precision of the results. The lower SE value in the scores after the intervention (0.149) suggests that the estimated average derived from the sample is closer to the true value.Coefficient of Variation (CoV) is a comparison of the standard deviation to the mean. The lower CoV value in the scores after the intervention (0.152) indicates that the relative variation in the data has decreased, meaning that the participants may have become more consistent in achieving results.Overall, this table illustrates that the intervention tends to improve participants' outcomes with less variation in results and better precision in estimating the mean.

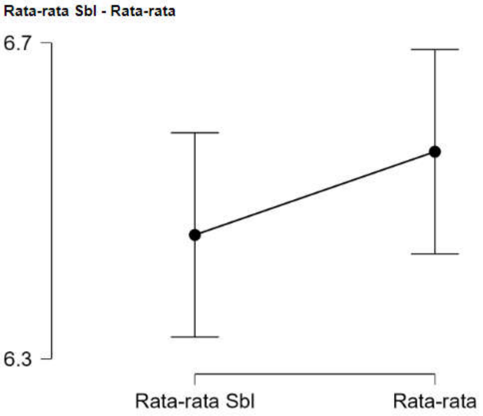

Figure 1. Descriptives of Mean-Measuring Plots

Based on figure 1 it is found that the graphically shows an improvement, but both the table and the descriptive chart are merely data and not statistical references as it follows.This is an illustration of the significant gap between the mean pre-test value and the mean post-test value. It illustrates perfectly the table 1.

Meanwhile, the results of the assumption test (Shapiro-Wilk) before the t-Test, needs to meet minimum requirement of assumptions (p > 0,05) as follows:

Table 2. Test of Normality (Shapiro-Wilk)

			W	P
Rata-rata Sbl	-	Rata-rata	0.963	0.164

Note. Significant results suggest a deviation from normalityby JASP 0.18.0.0.

Based on table 2 it is found that the Shapiro-Wilk normality test shows that the data is normally distributed (p > 0.05) and fulfills parametric test requirements.The Shapiro-Wilk normality test table indicates that the data is normally distributed. This is evidenced by the p-value being greater than 0.05, which suggests that the distribution of the data does not significantly deviate from a normal distribution. Therefore, the data fulfills the requirements for a parametric test. This is important as it validates the use of statistical techniques that assume a normal distribution, and it ensures the reliability and validity of the subsequent statistical analysis.

Then, here is the result of classic paired sample T-Test;

Table 3. Paired Samples T-Test

Measure 1	Measure 2	t	df	p	Mean Difference	SE Difference	Cohen's d	SE Cohen's d
Rata-rata sbl	Rata-rata	-1.161	44	0.252	-0.105	0.091	-0.173	0.082

Note. Student's t-test by JASP 0.18.0.0.

Based on table 3 it is found that the statistical results indicate that there is no significant difference between the "Average Before" and "Average Now" measurements. The t-statistic, which measures the difference between the two averages, is -1.161 with 44 degrees of freedom. The p-value, at 0.252, suggests that the observed difference is not statistically significant as it exceeds the typical significance threshold (e.g., 0.05). The mean difference between the two measurements is -0.105, indicating a slight decrease in the results in the "Average Now" measurement compared to the "Average Before." The Cohen's d value of -0.173 shows the magnitude of this difference in standard deviation units, with a negative value indicating that the "Now" measurement is lower than the "Before" measurement.

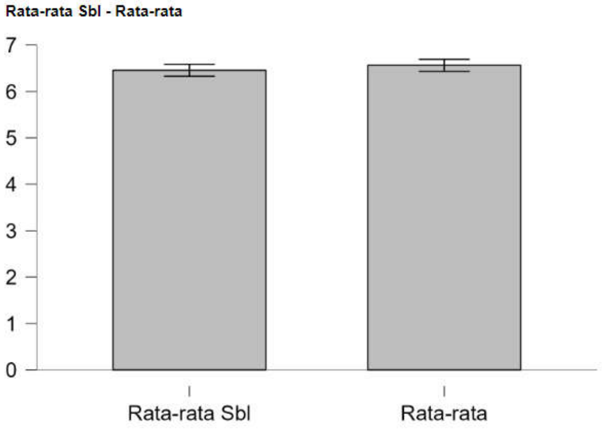

Figure 2. Bar Plots

As described in Table 3 and visually confirmed by the bar plot in Figure 2, the data findings are consistent. The bar plot provides a visual representation of the data, with the horizontal axis illustrating what is described in the table. The construction of the 95% confidence interval with a high level of confidence further supports these findings. This visual aid complements the tabular data, making it easier to understand the distribution and relationships within the data.

The result of Bayesian Paired T-Test as follows;

Table 4. Bayesian Paired Samples T-Test

Measure 1	Measure 2	BF_{10}	error %
Rata-rata Sbl	- Rata-rata	0.303	0.045

Note Bayesian factors by JASP 0.18.0.0.

Based on table 4 it is found that the result of the Bayesian Paired Sample T-Test indicates that Bayes Factor (BF_{10}) between Measure 1 (Mean Before) and Measure 2 (Mean) is 0.303 with an error rate of approximately 0.045%. This indicates a moderate level of evidence for the difference between the two measures, favoring the alternative hypothesis. In other words, this result suggests that the Bayesian method is more adaptive in measuring differences compared to the classical statistical method.

Furthermore, the results of the inferential calculations and posterior distribution for probability analysis are visualized in the diagram as follows:

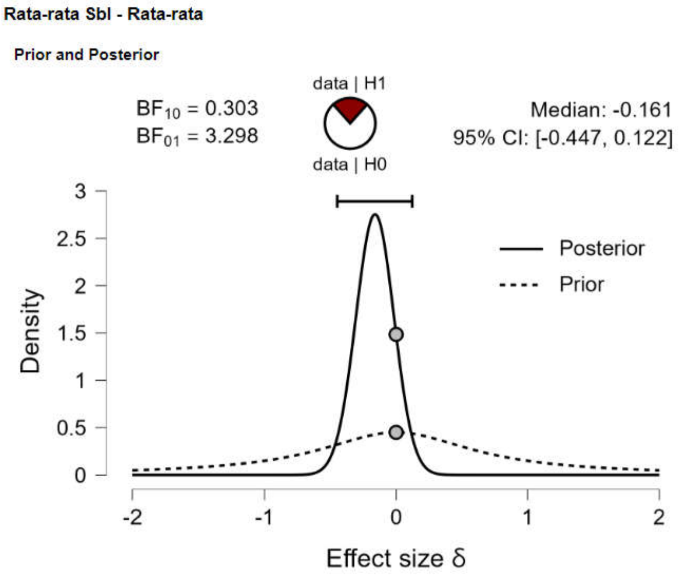

Figure 3. Inferential Plotsof Mean-Measuring Test

Based on figure 3 it is found that the provides a visual representation of the results obtained from the Bayesian Paired Sample T-Test, as detailed in the corresponding table. The graph indicates that the Bayes Factor (BF_{10}) between Measure 1 (Pre-Test) and Measure 2 (Post-Test) is 0.303. This value is a measure of the strength of evidence in favor of the alternative hypothesis, with values closer to 0 indicating stronger evidence for the null hypothesis. The error rate associated with this calculation is 0.045% (-0.447, 0.122), which represents the range within which the true value of the Bayes Factor is likely to fall with 95% confidence. This visual aid complements the tabular data, making it easier to understand the statistical relationships within the data.

In Bayesian factor analysis, there is also the concept of Bayes Factor Robustness Check to test the stability of findings, sensitivity of results, and potential outliers. Based on the data, it can be visualized as follows:

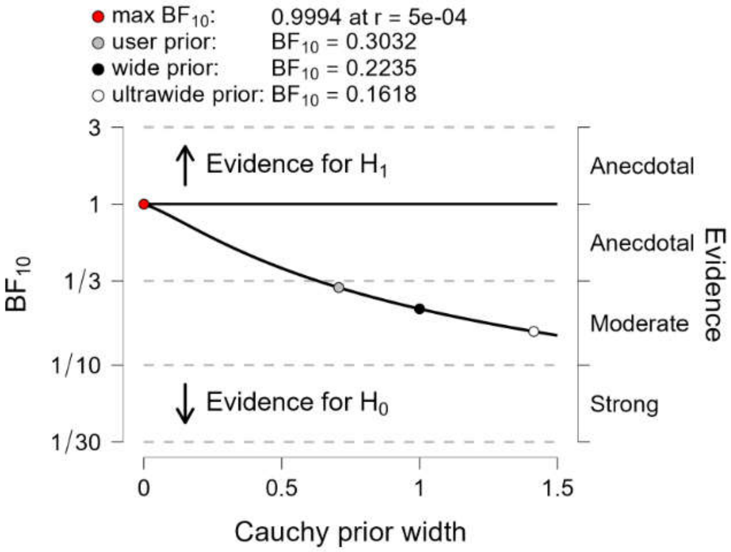

Figure 4. Bayes Factor Robustness Check

Based on figure 4 it is found that the presents of Bayes Factor effect size for the alternative hypothesis across different prior widths. The Bayes Factor does not decrease significantly far from 1 for all prior widths, but remains in the moderate range. This suggests that the evidence for the alternative hypothesis is not strong and remains moderate regardless of the prior width used. The maximum Bayes Factor (BF10) observed is 0.9994 at a prior width of r=5e-04. For the user-defined prior, the Bayes Factor is 0.3032, indicating moderate evidence for the alternative hypothesis. For the wide prior, the Bayes Factor is slightly lower at 0.2235, and for the ultra-wide prior, it is at its lowest at 0.1618. These values suggest that the strength of evidence for the alternative hypothesis decreases as the prior width increases.

As a significance control and for further decision-making reference, Bayesian sequential analysis used to test hypotheses by analyzing data sequentially. Here is sequential analysis that presented by the researchers:

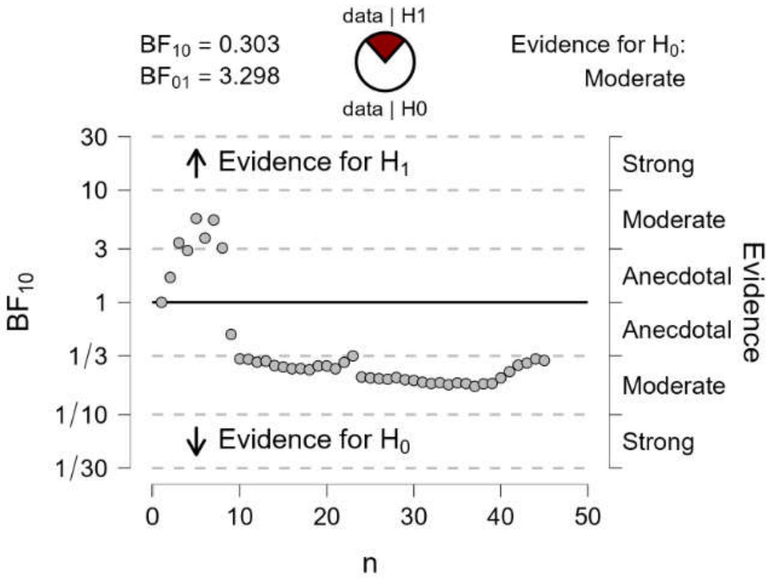

Figure 5. Sequential Analysis

Based on figure 5 it is found that the visualizes of hypotheses typically tested in Bayesian sequential analysis, which are the null hypothesis (H0) and the alternative hypothesis (H1). In this context, the null hypothesis (H0) posits that "Moderate" is the most likely or closest to the observed outcome based on the collected data. This means that, according to the null hypothesis, the data collected so far suggests that the "Moderate" category is the most probable. This hypothesis is tested against the alternative hypothesis (H1), which posits a different outcome. The process of Bayesian sequential analysis involves continuously updating our beliefs about these hypotheses as more data is collected, allowing us to make more accurate predictions over time.

As a result of the effectiveness test using a quantitative approach, descriptive data has been found in the descriptive table. Values such as t-statistic, degrees of freedom (df), and p-value (p) are often used to measure the effectiveness of a method or treatment. In this case, the t-value is -1.161, df is 44, and p is 0.252. The t-value measures how statistically significant the difference between two groups or conditions being compared is. In this case, the t-value is -1.161. This value indicates that the difference between the two groups or conditions being compared is not statistically significant. In other words, there is no statistically significant difference between the groups or conditions being compared. Degrees of freedom (df) refer to the amount of data used in the analysis. In this case, df is 44. The higher the df value, the more data is used in the analysis. This can increase the accuracy of the analysis and help detect smaller differences. However, in this case, df is relatively large, indicating that the analysis is based on a substantial amount of data. The p-value (p) is a measure of statistical significance. In this case, p is 0.252. A high p-value indicates that there is no statistical significance in the observed difference. In other words, the difference between the groups or conditions is not statistically significant.

In the context of effectiveness analysis, these results can be interpreted as follows: The method or treatment being tested is not statistically proven to be effective. The analysis does not support the presence of a significant difference between the groups receiving the treatment and those that do not. However, it's important to remember that statistical analysis is just one tool in assessing the effectiveness of a method or treatment. There are many other factors that can influence the results, and statistical analysis does not always reflect the actual impact of an action or method. In conclusion, the results of the analysis, with a t-value of -1.161, *df* of 44, and *p*-value of 0.252, indicate that the method or treatment being tested is not statistically proven to be effective. This analysis can serve as a starting point for further evaluation or changes in the approach used. This may be due to participants' doubts about the validity and accuracy of the chatbot used in various cases.

The results of the Bayesian paired t-test, which yielded a Bayes Factor (Bf_{10}) value of 0.303 and an error rate of around 0.045 percent, provide a deep understanding of the extent to which the observed data supports the tested hypothesis in the context of the analysis. Bayes Factor (Bf_{10}) is a measure used in Bayesian analysis to assess the strength of evidence supporting the alternative hypothesis (H_1) compared to the null hypothesis (H_0). A Bf_{10} value of 0.303 indicates that the existing evidence does not strongly support the alternative hypothesis compared to the null hypothesis. In this case, the alternative hypothesis may be that there is a significant effect or difference, while the null hypothesis is that there is no significant effect or difference. With a Bf_{10} value of 0.303, this suggests that the existing evidence is not strongly in favor of the alternative hypothesis. It may indicate that the observed data is not strong enough to support a significant difference or effect between the two groups or conditions being compared. The error rate of around

0.045 percent is the significance level used in Bayesian analysis. This error rate measures the acceptable level of error in making statistical decisions. In this context, an error rate of around 0.045 percent indicates that this analysis has a very high level of statistical significance, meaning that the findings are considered highly statistically significant. However, despite the very low error rate, the low Bf_{10} value indicates that the existing evidence still does not strongly support the alternative hypothesis. Therefore, the results of this analysis suggest that there is a possibility that the observed data may not be strong enough to support the claim of a significant effect or difference. In further interpretation, it is essential to consider the specific context of this analysis and the practical implications of these results. These findings can serve as a basis for further evaluation or advanced analysis and can help in making further decisions related to the tested hypothesis.

In the context of the analysis with a Bayes Factor (Bf_{10}) of 0.303 and an error rate of around 0.045 percent, the generated interval plot can illustrate the confidence interval for the observed parameter. This interval will encompass possible values for the parameter based on the observed data and the chosen confidence level. Interval plots typically depict two vertical lines representing the boundaries of the confidence interval. This interval can have a certain confidence level, such as 95 percent, which means that the parameter is estimated to be within that interval with a 95 percent confidence level. In other words, there is a 95 percent chance that the true parameter value falls within that interval.The mentioning of"95% CI: [-0.447, 0.112]" represents a 95 percent confidence interval for the observed parameter in the statistical analysis. This confidence interval provides an estimate of the range of possible values for the parameter with approximately 95 percent confidence. The lower bound of the confidence interval is -0.447, while the upper bound is 0.112. This means that based on the observed data and the statistical analysis conducted, we have about 95 percent confidence that the true value of the parameter lies within the range of -0.447 to 0.112. In other words, there is a 95 percent chance that the parameter has a value between -0.447 and 0.112. This confidence interval provides information about the level of uncertainty in the parameter estimate. The narrower the confidence interval, the higher the confidence in the parameter's value within the given range[26]. So, in the context of this data, we have about 95 percent confidence that the parameter value falls within the range of -0.447 to 0.112. This confidence interval provides valuable information about the extent of our confidence in the parameter estimate based on the observed data. In sequential evidence plots, founded that "evidence for H_0: Moderate" in Bayesian sequential analysis means that the data observed or collected is more in favor of the idea that "Moderate" is the most likely outcome or closest to the observed outcome based on that data. Finally, the discussion on the acceptability and effectiveness of using AI chatbots based on Large Language Models (LLMs) in Arabic language learning as a second language (L2) revealed several important insights that can enhance our understanding of how AI chatbots can influence the Arabic language learning process and their impact on learners.

The research highlights the significance of the context and learners' experiences in accepting and integrating AI chatbots into Arabic language learning. Qualitative findings indicate that the majority of learners are enthusiastic about using AI chatbots. They see them as valuable tools for improving their understanding and proficiency in the Arabic language. However, some learners also expressed concerns about the AI chatbots' ability to understand colloquial Arabic and dialects

[26] Johnny Van Doorn et al., *"The JASP Guidelines for Conducting and Reporting a Bayesian Analysis,"* Psychonomic Bulletin & Review 28, no. 3 (June 2021): 813–26. DOI: https://doi.org/10.3758/s13423-020-01798-5

commonly used in everyday conversations. This suggests that the context of using AI chatbots in Arabic language learning needs to be carefully considered to maximize their benefits.

These findings provide insights into the importance of learners' motivation and confidence in accepting and using AI chatbots. Learners with high motivation to learn Arabic and confidence in using technology are more likely to embrace AI chatbots as learning aids. Therefore, learners with lower motivation or less confidence in their technological abilities may require additional support to integrate AI chatbots into their learning. This research also underscores the need for validity in using AI chatbots in the context of Arabic language learning as mutual concern. The validity of using AI chatbots in teaching Arabic should be ensured to ensure that learners not only learn well but also have a good understanding of the material being taught. Therefore, the development of AI chatbots should consider accuracy in understanding Arabic language, not only in the context of daily conversations, but throughout various subject matters of Arabic learning. This aligns with the research by Hong and Fourney[27], although not on the same platform, and contradicts the findings of several researchers[28], for setting in different objects and categories and also in line with content accuracy concern.[29]

However, it is essential to note that this research has some limitations. Firstly, this research was conducted in a specific educational institution with a group of learners with varying levels of Arabic language proficiency. The results of the research may differ when applied to groups of learners with different characteristics. Moreover, the varying answers from chatbots' room to room responses to the same question with different definitions raised concerns, particularly in the context of Arabic literature. It may not have been expected to exhibit such behavior, especially in Arabic literature.Mostly in Arabic Language fileds, the responses should be more consistent and precise. Aspect of science such as technology, information, sociology, medic, and psychology may approach high accuracy with transformer-based analyses[30]. However, this might not be the case for Arabic literature, which has a strong foundation and strict rituals in its hierarchy of validation. Secondly, this research focused on the use of AI chatbots in Arabic language learning and may not be directly applicable to other language learning contexts or in different educational settings.

[27] Mike Perkins, *"Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond,"* Journal of University Teaching and Learning Practice, British University, Vietnam 20, no. 2 (February 22, 2023). DOI: https://doi.org/10.53761/1.20.02.07

[28] Ahlam Fuad and Maha Al-Yahya, *"Recent Developments in Arabic Conversational AI"* IEEE Access Volume: 10 (2022): 23842 - 23859. DOI: https://doi.org/10.1109/ACCESS.2022.3155521; Islamiyah Sulaeman, Syuhadak Syuhadak, and Insyirah Sulaeman, *"ChatGPT as a New Frontier in Arabic Education Technology,"* Al-Arabi: Jurnal Bahasa Arab dan Pengajarannya= Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: http://dx.doi.org/10.17977/um056v7i1p83-105; Erham Budi Wiranto and Sri Suwartini, *"Artificial Intelligence and Trustworthy Principles in Global Islamic Education,"* Ushuluddin International Conference (USICON) 6 (2022): 80. https://conference.uin-suka.ac.id/index.php/USICON/article/view/1252

[29] Matthew K Hong et al., "Planning for Natural Language Failures with the Ai Playbook," Journal Name. Vol. Issue. 2021: 1–11. DOI: https://doi.org/10.1145/3411764.3445735

[30] Oscar NE Kjell et al., "Natural Language Analyzed with AI-Based Transformers Predict Traditional Subjective Well-Being Measures Approaching the Theoretical Upper Limits in Accuracy," *Scientific Reports* 12, no. 1 (2022): 3918. DOI: https://dx.doi.org/10.1038/s41598-022-07520-w; Joshua James Hatherley, *"Limits of Trust in Medical AI,"* Journal of Medical Ethics 46, no. 7 (2020): 478–81. DOI: https://dx.doi.org/10.1136/medethics-2019-105935; Douglas Johnson et al., *"Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model,"* Research Square 2023. https://doi.org/10.21203/rs.3.rs-2566942/v1

Closing

In the progressive landscape of Arabic language education, learners' motivation and self-confidence wield substantial influence over the acceptance and effective utilization of AI chatbots. Those who harbor high levels of motivation to master Arabic and possess a robust confidence in technology tend to seamlessly integrate AI chatbots into their learning journey. However, safeguarding the validity of AI chatbots in Arabic language instruction emerges as a paramount concern. These chatbots must exhibit a profound understanding of Arabic encompassing a wide array of subject matters, extending beyond everyday conversations. While participants acknowledge this critical aspect, lingering doubts persist regarding the accuracy of the learning materials. These doubts cast a shadow on the full embracement of AI utilization within this context and play a pivotal role in shaping its adoption.

Turning to the test results, both classical and Bayesian paired t-tests unveil an inconsequential difference in Arabic language proficiency among learners post-AI chatbot use, resulting in the dismissal of the efficacy hypothesis (Ha). This implies that while some degree of enhancement in Arabic language proficiency is evident, it falls short of attaining statistical significance. The meager Bayes Factor (Bf10) value, resting at 0.303, signals that the existing body of evidence fails to robustly support a substantial distinction between the pre-test and post-test groups. This suggests that the impact of AI chatbot usage may not be as pronounced as initially envisioned, potentially influenced by the characteristics of the learners, the learning environment, or the measurement methods applied in this study. Consequently, an in-depth scrutiny of these outcomes, coupled with an exploration of their underlying influencers, becomes imperative. Moreover, effective harnessing of AI chatbots in the realm of learning necessitates continuity and the simultaneous consideration of an array of assessment factors. The researcher is cognizant of the constraints imposed by time limitations and the confined research timeline, which encompassed only a few sessions. Nonetheless, AI retains its stature as a valuable component in the pursuit of Arabic language proficiency. Future research endeavors are poised to make significant contributions to scholarly discourse, with a particular focus on the accuracy of Arabic language learning materials and their adaptability across diverse educational tiers. In summary, this study underscores the importance of learners' motivation and self-assurance in embracing AI chatbots in Arabic language education. It also highlights the need for ongoing efforts to enhance the validity and accuracy of AI chatbots and their role in fostering language proficiency.

Acknowledgment

We would like to express our gratitude to STAI Al-Anwar and all the participants of final year's academic program for their contributions and support. Without this participation and cooperation, this research would not have been possible. Thank you for the time and effort you dedicated to this research. We hope our collaboration continues to grow and brings even greater benefits in the future.

Bibliografi

Abdulkader, Zena, and Yousra Al-Irhayim. *"A Review of Arabic Intelligent Chatbots: Developments and Challenges."* Al-Rafidain Engineering Journal (AREJ) 27, no. 2 (September 1, 2022): 178–89. DOI: https://doi.org/10.33899/rengj.2022.132550.1148

Anwar, Saiful, Guntur Cahaya Kesuma, and Koderi. *"Development of Al-Qawaid an-Nahwiyah Learning Module Based on Qiyasiyah Method for Arabic Language Education Department Students | Pengembangan Modul Pembelajaran al-Qawaid an-NahwiyahBerbasis Metode QiyasiyahUntukMahasiswaJurusan Pendidikan Bahasa Arab."* Mantiqu Tayr: Journal of Arabic Language 3, no. 1 (January 2, 2023): 11–24. DOI: https://doi.org/10.25217/mantiqutayr.v3i1.2830

Attaqiana, Mita Rosyda, Saptorini Saptorini, and Achmad Binadja. *"Pengembangan Media Permainan Truth and Dare Bervisi Sets Guna Memotivasi Belajar Siswa."* Jurnal Inovasi Pendidikan Kimia 10, no. 2 (2016): 1798 – 1806. DOI: https://doi.org/10.15294/jipk.v10i2.9533

Azmi, Aqil M, Abdulaziz O Al-Qabbany, and Amir Hussain. *"Computational and Natural Language Processing Based Studies of Hadith Literature: A Survey."* Artificial Intelligence Review 52 (2019): 1369–1414. DOI: https://doi.org/10.1007/s10462-019-09692-w

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, and Ruth Fong. *"Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims."* arXiv Preprint arXiv:2004.07213, 2020. DOI: https://doi.org/10.48550/arXiv.2004.07213

Chan, Cecilia Ka Yuk. *Assessment for Experiential Learning.* (1st ed.). Routledge London 29 September (2022): 378. DOI: https://doi.org/10.4324/9781003018391

Chiu, Thomas KF, Benjamin Luke Moorhouse, Ching Sing Chai, and Murod Ismailov. *"Teacher Support and Student Motivation to Learn with Artificial Intelligence (AI) Based Chatbot."* Interactive Learning Environments, 2023: 1–17. DOI: https://doi.org/10.1080/10494820.2023.2172044

De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. *"ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health."* Frontiers in Public Health 11 (April 25, 2023): 1166120. DOI: https://doi.org/10.3389/fpubh.2023.1166120

Fiantika, Feny Rita, Mohammad Wasil, Sri Jumiyati, Leli Honesti, Sri Wahyuni, Erland Mouw, Imam Mashudi, Nur Hasanah, Anita Maharani, and Kusmayra Ambarwati. *Metodologi Penelitian Kualitatif.* Padang: Get Press, 2022.

Fitriani, Fitriani, Muhammad Akmansyah, Ahmad Basyori, Erlina Erlina, & Koderi Koderi. *"Manajemen Pembelajaran Bahasa Arab di SMP Qur'an Darul Fattah (SQDF) Bandar Lampung."* Al Maghazi : Arabic Language in Higher Education, 1.2 (2023): 47-60. DOI: http://dx.doi.org/10.51278/al.v1i2.786

Fuad, Ahlam, and Maha Al-Yahya. *"Recent Developments in Arabic Conversational AI: A Literature Review."* IEEE Access Volume: 10 (2022): 23842 - 23859. DOI: https://doi.org/10.1109/ACCESS.2022.3155521

Gilbert, Stephen, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. *"Large Language Model AI Chatbots Require Approval as Medical Devices."* Nature Medicine 29, no. 10 (October 2023): 2396–98. DOI: https://doi.org/10.1038/s41591-023-02412-6

Hatherley, Joshua James. *"Limits of Trust in Medical AI."* Journal of Medical Ethics 46, no. 7 (2020): 478–81. DOI: https://dx.doi.org/10.1136/medethics-2019-105935

Hidayatullah, Achmad, and Csaba Csíkos. *"The Role of Students' Beliefs, Parents' Educational Level, and The Mediating Role of Attitude and Motivation in Students' Mathematics Achievement."* The Asia-Pacific Education Researcher, March 30, 2023. DOI: https://doi.org/10.1007/s40299-023-00724-2

Hong, Matthew K, Adam Fourney, Derek DeBellis, and Saleema Amershi. *"Planning for Natural Language Failures with the Ai Playbook,"* CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021: 1-11. DOI: https://doi.org/10.1145/3411764.3445735

Johnson, Douglas, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, and Eiman Jahangir. *"Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model,"* Research Square 2023. DOI: https://doi.org/10.21203/rs.3.rs-2566942/v1

Kjell, Oscar NE, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. *"Natural Language Analyzed with AI-Based Transformers Predict Traditional Subjective Well-Being Measures Approaching the Theoretical Upper Limits in Accuracy."* Scientific Reports 1 12, (2022): 3918. DOI: https://doi.org/10.1038/s41598-022-07520-w

Liang, Wei, and Hongsheng Dai. *"Bayesian Inference."* In Quantum Chemistry in the Age of Machine Learning, 233–50. Elsevier, 2023. DOI: https://doi.org/10.1016/B978-0-323-90049-2.00005-6

Mohamed, Amr M. *"Exploring the Potential of an AI-Based Chatbot (ChatGPT) in Enhancing English as a Foreign Language (EFL) Teaching: Perceptions of EFL Faculty Members."* Education and Information Technologies, 2023: 1–23. DOI: https://dx.doi.org/10.1007/s10639-023-11917-z

Munir, Munir. *"Pendekatan Struktural dalam Pelajaran Bahasa Arab."* Shaut al Arabiyyah 6, no. 1 (August 24, 2018): 13. DOI: https://doi.org/10.24252/saa.v6i1.5644

Mustaqim, Mustaqim. *"Metode Penelitian Gabungan Kuantitatif Kualitatif/Mixed Methods Suatu Pendekatan Alternatif."* Intelegensia: Jurnal Pendidikan Islam 4, no. 1 (2016): 1-9. DOI: https://doi.org/10.34001/intelegensia.v6i1.1351

Nita Amelia, (second) Noval Mulya Dava S., and (third) Muhammad Minan Chusni, *"Pemanfaatan Artificial Intelligence dalam Pembelajaran di Masa Pandemi| Prosiding Amal Insani Foundation,"* September 6, 2023. https://prosiding.amalinsani.org/index.php/semnas/article/view/10

Perkins, Mike. *"Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond."* Journal of University Teaching and Learning Practice, British University, Vietnam 20, no. 2 (February 22, 2023). DOI: https://doi.org/10.53761/1.20.02.07

Rahmah, Risa, Azizatul Muzdalifah, & Mu'alim Wijaya. *"Penggunaan Thariqah Mubasyarah sebagai Pembelajaran Bahasa Arab yang Efektif."* Al Maghazi : Arabic Language in Higher Education, 1.1 (2023): 23-29. DOI: https://doi.org/10.51278/al.v1i1.706

Ritonga, Mahyudin, Rizka Widayanti, Fitri Alrasi, Julhadi, and Syaflin Halim. *"Analysis of Arabic Language Learning at Higher Education Institutions with Multi-Religion Students."* Universal Journal of Educational Research 8 (September 1, 2020): 4333–39. DOI: https://doi.org/10.13189/ujer.2020.080960

Rumaisa, Fitrah, Yan Puspitarani, Ai Rosita, Azizah Zakiah, and Sriyani Violina. *"Penerapan Natural Language Processing (NLP) di Bidang Pendidikan."* Jurnal Inovasi Masyarakat 1, no. 3 (2021): 232–35. DOI: https://doi.org/10.33197/jim.vol1.iss3.2021.799

Shao, Sicong, Saleem Alharir, Salim Hariri, Pratik Satam, Sonia Shiri, and Abdessamad Mbarki. "*AI-Based Arabic Language and Speech Tutor."* In 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), 1–8. Abu Dhabi, United Arab Emirates: IEEE, 2022. DOI: https://doi.org/10.1109/AICCSA56895.2022.10017924

Shneiderman, Ben. *"Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy."* International Journal of Human–Computer Interaction 36, no. 6 (April 2, 2020): 495–504. DOI: https://doi.org/10.1080/10447318.2020.1741118

Sulaeman, Islamiyah, Syuhadak Syuhadak, and Insyirah Sulaeman. *"ChatGPT as a New Frontier in Arabic Education Technology."* Al-Arabi: Jurnal Bahasa Arab Dan Pengajarannya= Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: https://dx.doi.org/10.17977/um056v7i1p83-105

Van Doorn, Johnny, Don Van Den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, et al. *"The JASP Guidelines for Conducting and Reporting a Bayesian Analysis."* Psychonomic Bulletin & Review 28, no. 3 (June 2021): 813–26. DOI: https://doi.org/10.3758/s13423-020-01798-5

Wekke, Ismail Suardi, and Maimun Aqsha Lubis. *"A Multicultural Approach in Arabic Language Teaching: Creating Equality at Indonesian Pesantren Classroom Life."* Sosiohumanika 1, no. 2 (2008): 296-310. DOI: https://doi.org/10.2121/sosiohumanika.v1i2.337

Wiranto, Erham Budi, and Sri Suwartini. *"Artificial Intelligence and Trustworthy Principles in Global Islamic Education."* Ushuluddin International Conference (USICON) 6 (2022): 64–87. https://vicon.uin-suka.ac.id/index.php/USICON/article/view/1252