



Leveraging AI-Generated Images for Visual Stimulus-Based Arabic Speaking Assessment

Rovi Udin¹, Riska Hangkiho², Muhammad Jundi³, Renni Hasibuan⁴

¹Madrasah Tsanawiyah Alkhairaat Dudewulo

²Institut Agama Islam Negeri Sultan Amai Gorontalo

^{3,4}Universitas Islam Negeri Maulana Malik Ibrahim Malang

<rofiudinpba@gmail.com> <ikahangkiho@gmail.com>

<rennihasibuan88@gmail.com> <jundijundi10@gmail.com>

Abstrak: Dalam beberapa tahun terakhir, integrasi teknologi kecerdasan buatan (AI) telah merevolusi berbagai aspek pendidikan, termasuk pembelajaran bahasa dan penilaiannya. Studi ini bertujuan untuk mengeksplorasi pemanfaatan gambar yang dihasilkan oleh AI dalam pembelajaran bahasa dan aplikasinya dalam asesmen keterampilan berbicara bahasa Arab. Penelitian ini mengkaji proses pembuatan gambar untuk tujuan asesmen keterampilan berbicara bahasa Arab menggunakan platform AI. Dengan analisis terhadap berbagai sumber literatur terkait, pendekatan kualitatif digunakan untuk memahami bagaimana gambar AI-generated dapat digunakan dalam asesmen keterampilan berbicara bahasa Arab yang menggunakan stimulus visual. Hasil penelitian menunjukkan bahwa terdapat dua jenis asesmen keterampilan berbicara yang menggunakan stimulus visual, yaitu: menjelaskan gambar dan menjawab pertanyaan secara lisan berdasarkan gambar. Proses pengembangan instrumen penilaian melibatkan analisis kompetensi, penentuan indikator capaian, formulasi kisi-kisi, serta penyusunan item dan skala penilaian. Proses persiapan gambar dengan bantuan AI dilaksanakan pada proses penyusunan item. Platform kecerdasan buatan yang digunakan adalah Bing AI Image Generator karena aksesibilitasnya yang mudah dan gratis. Penting untuk dipahami bahwa keberhasilan dalam memproduksi gambar yang sesuai tujuan asesmen sangat bergantung pada formulasi dan kelengkapan prompt yang dituliskan. Hasil ini menekankan pentingnya peran prompt yang dirancang dengan baik oleh penyusun instrumen asesmen. Prompt yang baik dapat memaksimalkan efektivitas gambar yang dihasilkan oleh AI dalam penilaian keterampilan berbicara berbasis stimulus visual. Hal ini pada gilirannya memberikan informasi yang berharga untuk perancangan pembelajaran dan pengembangan asesmen dalam konteks pembelajaran bahasa.

Kata Kunci: *Kecerdasan Buatan (AI); Pembelajaran Bahasa; Keterampilan Berbicara Bahasa Arab; Stimulus Visual; Instrumen Penilaian*

Abstract: In recent years, the integration of artificial intelligence (AI) technology has revolutionized various aspects of education, including language learning and assessment. This study aims to explore the utilization of AI-generated images in language learning and its application in the assessment of Arabic speaking skills. The research examines the process of creating images for the purpose of assessing Arabic speaking skills using AI platforms. By analyzing various related literature sources, a qualitative approach is employed to understand how AI-generated images can be used in the assessment of Arabic speaking skills using visual stimuli. The results indicate that there are two types of speaking skill assessments using visual stimuli: describing images and responding to questions orally based on images. The development process of the assessment instruments involves competence analysis,

determination of achievement indicators, formulation of assessment grids, and the creation of items and rating scales. The image preparation process with AI assistance is carried out during the item creation phase. The AI platform used is the Bing AI Image Generator due to its easy accessibility and free availability. It is essential to understand that the success of producing images suitable for assessment purposes heavily depends on the formulation and completeness of the written prompt. These findings underscore the importance of well-designed prompts by the assessment instrument developers. A well-crafted prompt can maximize the effectiveness of AI-generated images in visual stimulus-based speaking skill assessments. This, in turn, provides valuable insights for instructional design and assessment development in the context of language learning.

Keywords: *Artificial Intelligence (AI); Language Learning; Arabic Speaking Skills; Visual Stimuli; Assessment Instruments*

1. INTRODUCTION

Assessment stands as a cornerstone in education, serving the vital purpose of gauging students' mastery of targeted educational goals and their proficiency levels (Jundi & Ali, 2023; Triantafillou et al., 2008; 2022, علييمات). Its significance lies in its role in guiding instructional enhancements and pinpointing areas necessitating pedagogical refinement (Mardapi, 2016; 2010, عواضة). Within formal educational frameworks, assessment typically takes the form of periodic evaluations, functioning as a pivotal tool for tracking students' learning progress (Hasibuan & Jundi, 2023; علام, 2005; عودة, 2005). Consequently, educators cannot afford to overlook the pivotal role of assessment in the teaching-learning continuum (Arikunto, 2009; Kementrian Pendidikan dan Kebudayaan, n.d.). While the educator's primary duty encompasses instruction, the assessment of students' competencies equally holds paramount importance. Hence, assessment emerges as an indispensable component in the educational landscape, facilitating the enhancement of educational quality and the attainment of specific learning objectives (Adam et al., 2023; Fuentealba, 2011; Gipps, 1998). This pertinence extends to the realm of speaking skills assessment, where gauging communicative proficiency is equally imperative for facilitating comprehensive language development and achieving targeted learning outcomes.

Assessing Arabic speaking skills (maharah kalam) can be done through various types of assessments, including Oral Reading, Picture Description, Retelling, Free Speech, Interview, and Discussion (Asrori et al., 2019). However, in practice, the use of visual stimuli such as pictures often poses a challenge for Arabic language educators and instructors. This is due to limited access to images that align with the required criteria and learning indicators. Another challenge arises when detailed images are needed for instructional or evaluative purposes but are difficult to find for free on the internet. For example, difficulties may arise in finding fine quality pictures of children wearing Muslim attire, caps, or hijabs as Arabic curriculum required (Ali & Jundi, 2023). Additionally, images depicting specific activities with figures dressed modestly are also required. Therefore, having flexible and tailored image sources that meet the needs of teaching and evaluation becomes crucial. This research will focus on types of assessments for Arabic speaking skills that utilize visual stimuli, such as describing pictures and answering questions orally based on images.

In recent years, the integration of artificial intelligence (AI) technologies has profoundly influenced various domains, including education and language learning (Ji et al., 2022; Kshirsagar et al., 2022; Yanhua, 2020; Yin, 2021). With the advancement of AI, particularly in the realm of image generation, new opportunities arise for enhancing language

assessment methodologies. Language assessment, particularly in speaking proficiency evaluation, has traditionally relied on conventional methods (Bernstein, 2006; Rosser, 1979), often facing challenges related to providing images used for assessment (Plough, 2018). In this context, the utilization of AI-generated images presents a promising avenue for overcoming these challenges and advancing language assessment practices.

Despite the potential benefits of AI-generated images in language assessment, research exploring their utilization in the context of Arabic language speaking proficiency evaluation remains limited. Arabic language learning presents unique linguistic and cultural characteristics that warrant tailored assessment approaches. Therefore, investigating the feasibility and efficacy of integrating AI-generated images into Arabic speaking assessment is crucial for advancing language assessment practices in Arabic language education. Moreover, exploring the challenges and obstacles encountered during the image creation process is essential for identifying potential limitations and areas for improvement. By addressing these challenges, researchers and educators can refine the utilization of AI-generated images in language assessment and maximize their effectiveness in evaluating Arabic language speaking proficiency.

Furthermore, understanding the theoretical foundations of language assessment and the principles underlying AI image generation is imperative for ensuring the validity and reliability of assessment outcomes (Latief, 2016; Park et al., 2015). By examining the process of AI image creation alongside established theories of language assessment instrument development, this study aims to illuminate the intersection between technological advancements and theoretical constructs in language assessment. Through this correlation, the research endeavors to uncover how innovative technologies, such as AI-generated images, can be

harmonized with existing theoretical frameworks to enhance the efficacy and validity of language assessment practices. By drawing parallels between the intricacies of AI image generation and the principles underlying language assessment instrument development, this investigation seeks to offer valuable insights into how technological innovations can be effectively integrated into the assessment landscape, ultimately contributing to the evolution and refinement of language assessment methodologies.

Previous studies have delved into various aspects of AI integration in education, offering valuable insights into its potential benefits and challenges. For instance, Melati et al. (2023) explored the use of AI and technology in creating English learning materials aligned with the Merdeka Belajar Learning Guidelines, highlighting its positive impact on enhancing learning quality. Mukti (2023) emphasized the positive influence of AI integration in elementary education, particularly in enriching learning experiences. Additionally, Rina (2023) investigated AI's role in project-based learning, revealing students' satisfaction with AI-enabled learning activities. However, challenges in education arise from AI implementation, as discussed by Rosyadi et al. (2023) and Zhai et al. (2021), including issues related to inappropriate AI usage, shifting teacher and student roles, and ethical considerations. Furthermore, Prananta et al. (2023) explored ChatGPT's utility as an information search and scientific communication tool, while Wibisono and Pramudita (2023) demonstrated the suitability of AI-based learning media for mathematics education in elementary schools. Rahman et al. (2023) also found significant effects of AI-based blended learning models on students' science literacy.

Despite the wealth of research on AI integration in education, there remains a research gap concerning its application in Arabic language speaking assessment. The

existing literature primarily focuses on AI's utilization in various subjects and aspects of education, such as English learning, mathematics education, and science literacy. However, limited attention has been given to exploring the potential of AI-generated images in enhancing Arabic language speaking assessment. Therefore, this study aims to fill this gap by investigating the effectiveness and feasibility of utilizing AI-generated images for assessing Arabic language speaking proficiency. Through a comprehensive analysis, this research seeks to shed light on the innovative use of AI technology in language assessment and contribute to the advancement of Arabic language education.

Overall, the proposed research aims to contribute to the advancement of language assessment practices by investigating the integration of AI-generated images into Arabic language speaking assessment. The research problem addressed in this study revolves around the development and implementation of an assessment instrument for Arabic language speaking proficiency utilizing AI-generated images. Specifically, the inquiry focuses on investigating the process of creating and refining such assessment tools, aiming to explore their efficiency and effectiveness in evaluating students' speaking skills. By bridging the gap between technological innovations and language assessment methodologies, this research seeks to enhance the efficiency, objectivity, and validity of Arabic language assessment practices, ultimately benefiting language learners and educators alike..

2. METHOD

In this study, a qualitative approach will be employed to investigate the utilization of AI-generated images in language assessment. The research methodology will encompass two primary methods: documentation in the literature review and observation of open access AI image generation platforms to assess their efficiency (Baker, 2006). Through a

comprehensive review of related literature (Cowie, 2009), relevant studies will be analyzed to glean insights into the efficacy and applications of AI-generated images in language assessment contexts. This documentation method will involve scrutinizing existing literature, academic papers, and reports to ascertain the current state of research, identify trends, and evaluate methodologies and outcomes (Bowen, 2009).

Concurrently, observational methods will be utilized to observe and analyze the process of AI image generation across different platforms. This observation will entail firsthand exploration and experimentation with AI image generation tools, allowing for an in-depth understanding of their functionalities, capabilities, and limitations.

The collected research data will undergo analysis employing an interactive data analysis approach, which involves ongoing analysis throughout the data collection process to guide research focus and refine analytical strategies (Rukajat, 2018). Qualitative analysis techniques, particularly content analysis, will be utilized to interpret and extract insights from the documented data (Miles, 2014). Through content analysis, researchers will systematically examine and interpret the information gathered from literature sources, identifying key themes, patterns, and findings relevant to the research objectives. This methodological approach aims to provide a rigorous and comprehensive analysis of the data, facilitating the generation of meaningful interpretations and conclusions regarding the utilization of AI-generated images in language assessment.

3. FINDINGS AND DISCUSSION

Assessing Speaking Proficiency Using Visual Stimuli

The assessment methods for Arabic speaking skills encompass pronunciation tests, oral sentence structure assessments, tests utilizing visual stimuli, and tests

gauging expression delivery (Mahmudi & Nurbaha, 2023). Speaking assessments should be conducted through oral practice, either directly or through monologues by creating videos (Dinata et al., 2021; Hosna et al., 2023). This research focuses on types of assessments for Arabic speaking skills that utilize visual stimuli, such as describing pictures and answering questions orally based on images. These two assessment modalities enable educators to explore students' speaking abilities with a focus on specific competencies and achievements. By referring to prepared visual stimuli, desired accomplishments can be measured precisely. High-order thinking skills (HOTS) assessment for the four Arabic language skills can be effectively executed by ensuring that the assessment format aligns with the dimensions of HOTS thinking processes, namely creating, evaluating, and analyzing (Hanief, 2022).

This integration allows for a comprehensive evaluation of students' speaking abilities while emphasizing critical thinking and higher-order cognitive processes.

The assessment procedure utilizing visual stimuli is conducted in accordance with standard speaking assessment protocols as shown in figure 1. Prior achievements are formulated and delineated into indicator matrices, serving as the basis for the construction of assessment items. Throughout this process, educators also prepare the necessary images for evaluating students' speaking abilities. The images prepared are AI-generated visuals crafted by educators utilizing an AI image generation platform, namely the Bing Image Generate tool. Additionally, teachers develop a grading rubric to ensure a more objective assessment.

Figure 1. Assessment instrument development process



In general, the steps involved in the development of a high-quality Arabic language test that accurately measures proficiency are outlined as follows (Roviin, 2018). Firstly, a thorough examination of the curriculum or syllabus is conducted, encompassing fundamental competencies, indicators, and key/sub-topics to serve as the foundation and reference for test construction. Next, a test blueprint is devised, including outlining core competencies and indicators, specifying main and sub-discussion topics, determining the number and format of test items. Subsequently, test item construction and scoring guidelines are developed. The items undergo scrutiny to rectify errors such as numbering and wording inconsistencies. Following this, a pilot test is administered to gauge the effectiveness and appropriateness of the test. The test items along with their corresponding answers are then analyzed utilizing validity

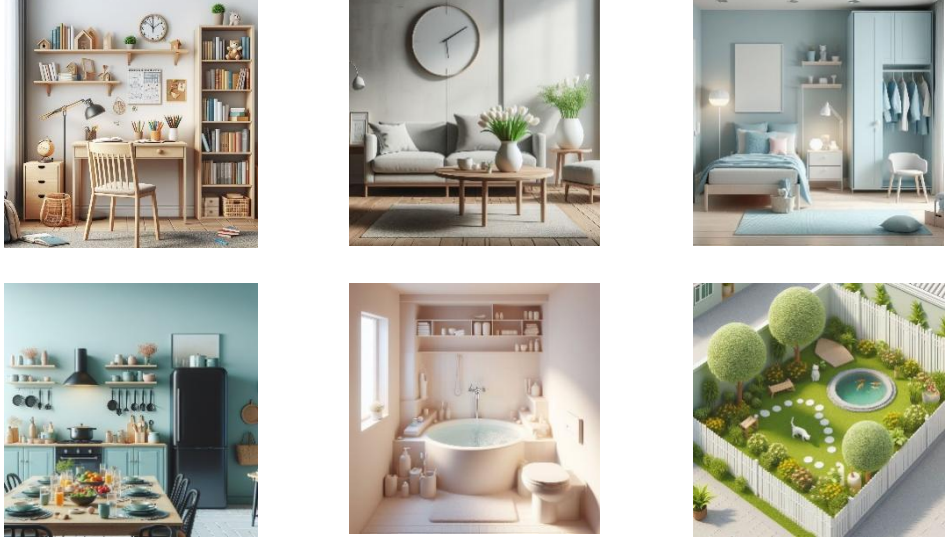
and reliability testing formulas, and difficulty level assessment. Any necessary revisions to the test items are made based on the analysis results. Finally, the test is administered, and the results are interpreted to assess proficiency levels.

According to Cherice Montgomery, speaking proficiency encompasses six key aspects: pronunciation, fluency, vocabulary, content, accuracy, and comprehension, as well as strategic competence (Matrokhim, 2021). Other perspectives suggest that the components evaluated in speaking skills assessment encompass: pronunciation, grammar, vocabulary, fluency, speech content, and comprehension (Widodo & Qosim, 2021). In the assessment table provided, there are five aspects being evaluated, namely: Pronunciation Clarity, Grammar Mastery, Vocabulary Proficiency, Fluency, and Comprehension, as can be seen in Figure 2.

Students' scores are determined on a scale of 1 to 5 for each evaluated aspect. In terms of pronunciation proficiency, the highest score is awarded if students exhibit no speech defects, while the lowest score is given if there are numerous defects hindering speech comprehension. Regarding grammatical proficiency, high scores are granted to students capable of communicating with well-structured sentences, whereas low scores are assigned if there are multiple errors impeding sentence coherence. Concerning vocabulary usage, the highest score is bestowed upon students adept at employing a rich vocabulary or idiomatic expressions, whereas the lowest score is attributed to those with limited vocabulary

hindering smooth communication. This aspect warrants further attention as vocabulary constitutes a crucial element in acquiring a second language (Jundi, 2023b; Jundi & Hasibuan, 2023). In terms of fluency, the highest score is granted if students speak fluently akin to native speakers, while the lowest score is given if speech is disjointed or halting. Regarding comprehension, the highest score is awarded if students effortlessly understand all information, whereas the lowest score is assigned if they struggle to comprehend spoken content. The latter two aspects are particularly relevant to assessing speaking proficiency through interviews (Asrori et al., 2019).

Figure 2. Assessment instruments and grading rubrics.

	
وصف الصورة بالكلام!	1- ما هذا؟ 2- ماذا فيها؟ -ماذا تفعل فيها؟

No	Name	Scoring Aspects					Total
		Speech	Grammar	Vocabulary	Fluency	Comprehension	
1							
2							
3							
....							

Assessing speaking skills, or maharatul kalam, represents a critical aspect of language testing. Harris contends that among language skills, speaking proficiency is notably challenging to evaluate (Asrori et al., 2019). Speaking proficiency requires the concurrent utilization of multiple abilities, encompassing pronunciation, which includes vocal and consonantal characteristics, stress patterns, and intonation; grammar; vocabulary or mufradat; fluency in articulating or pronouncing expressions; and comprehension, indicating the ability to grasp ideas conveyed by interlocutors (Ainin, 2016; Dalle, 2019). It is paramount to underscore that assessing students' speaking proficiency is ideally conducted through oral tests rather than written ones. Oral tests prove significantly more effective in gauging individuals' comprehension and capability in speaking skills or maharatul kalam (Miladya, 2015).

Several issues have been identified regarding the assessment of Arabic speaking skills among primary level students, including the following: 1) Teachers frequently overlook the assessment of speaking skills due to an overemphasis on written evaluations, 2) Insufficient time allocation also relegates speaking skill evaluation to a secondary priority for teachers, 3) Teachers acknowledge the challenge of developing appropriate assessment instruments for evaluating speaking skills, and 4) Limited student proficiency hampers the continuous assessment of speaking skills (Kholisoh, 2018). Furthermore, similar to other forms of assessment, analysis of the speaking assessment instrument is also crucial (Yasin & Tarauni, 2023). This ensures the quality of the assessment and ensures that it accurately measures what is intended to be measured (Jundi, 2023a, 2023c). Consequently, evaluation can proceed effectively.

AI-Generated Image as Arabic Speaking Assessment Media

As previously elucidated, there is no substantial difference in the implementation of speaking assessments using images generated by artificial intelligence (AI) compared to general speaking assessment procedures. The form of assessment can take any technique and method (Zaharo, Chamidah, et al., 2024). The primary distinction lies in the utilization of images generated by AI as visual stimuli. Therefore, the focus of discussion will be directed towards methods to generate relevant images in accordance with assessment needs.

Presently, amidst the advanced progress of artificial intelligence, there exist various methods to access and produce AI-generated images tailored to specific requirements. Within the scope of this research, the chosen artificial intelligence platform is the Bing AI Image Generator. Developed by Microsoft, the Bing AI Image Generator is a system utilizing artificial intelligence technology to generate new images based on user requests (Agarwal et al., 2023; Aiumtrakul et al., 2023; Al-Ashwal et al., 2023; Zaharo, Hangkiho, et al., 2024). One of the key advantages considered in selecting this AI platform is the provision of free access by Bing. This affords researchers or other users the flexibility to access and utilize the service without the need for subscription fees or additional licensing costs.

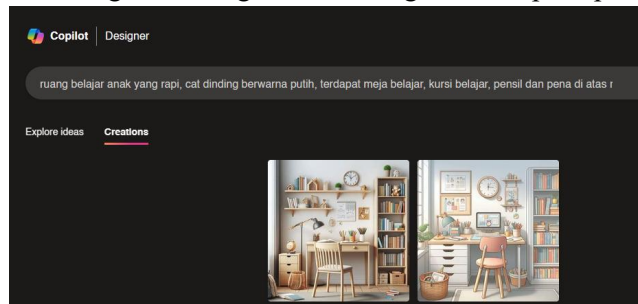
There is a plethora of AI platforms available for various purposes. Among the most commonly encountered AI platforms are ChatGPT and Bing AI. Both of them are widely recognized and offer free access to users. The research suggests that Bing and ChatGPT can be utilized to generate valid and challenging reasoning-based multiple-choice questions, albeit improvements are still necessary (Agarwal et al., 2023). In another study, Bing AI has shown promise in augmenting data extraction in systematic reviews through the validation and enhancement of extracted data items (Hill et al., 2024).

Indeed, Bing AI exhibits superior accuracy and specificity surpassing Google's Bard, ChatGPT-3.5, and ChatGPT-4 (Al-Ashwal et al., 2023). While AI can assist in literature search, its usage requires careful evaluation to ensure accuracy (Aiumtrakul et al., 2023).

In producing AI-generated images, a prompt is required to generate the desired images. With Bing AI, prompts can be written in various languages, but for this purpose, prompts in Bahasa Indonesia are utilized, as can be seen in Figure 3. There are no specific guidelines regarding the prompts used; users simply need to

describe the desired image as accurately as possible. The AI will then generate the requested image according to the prompt provided. If the generated image is not satisfactory, the prompt can be adjusted accordingly. The probability of successfully generating images that meet one's desires heavily relies on the manner of crafting prompts and the completeness of the prompts themselves. As AI users, individuals must engage in ample practice and exercise their creativity in formulating prompts as creatively and comprehensively as possible.

Figure 3. AI-generated Image and the prompt.



Artificial intelligence (AI) has evolved significantly, transitioning from a rudimentary concept to a sophisticated and adaptable system capable of achieving diverse goals and tasks (Haenlein & Kaplan, 2019). This evolution prompts a need for deeper critical reflection on the challenges, risks, and ethical considerations associated with AI's integration into educational contexts (Zawacki-Richter et al., 2019). Moreover, AI holds promise in revolutionizing digital media education, with anticipated enhancements in teaching and learning quality and effectiveness (W. Wang & Liu, 2021). As AI continues to advance, automation and datafication are expected to see further development within the realm of education (Williamson et al., 2023).

The emergence of AI technologies signals a new era, offering opportunities to transform and augment the teaching-learning paradigm within educational

environments (Rosyadi et al., 2023). By harnessing AI, interactive and adaptive learning media can be developed to enrich students' learning experiences (Handayani et al., 2022; Wibisono & Pramudita, 2023). Furthermore, AI can facilitate the development of English learning media and materials that align with curriculum guidelines (Melati et al., 2023). In the realm of language modeling, AI can serve as a potent tool, exemplified by platforms such as ChatGPT (Prananta et al., 2023). Overall, AI stands as a valuable instrument in supporting the transformation of digital education and fostering better educational outcomes in the digital era (Mukti, 2023).

In the realm of education, the integration of artificial intelligence (AI) and machine learning is envisioned to revolutionize e-learning methodologies. Scholars anticipate that in the future, e-learning platforms will leverage AI to autonomously assess students' abilities, identify their weaknesses, discern their

talents, and analyze their interests (Buaton et al., 2022). This wealth of data can then be harnessed to tailor learning plans that cater to the individual needs and capabilities of students. Furthermore, deliberate and strategic integration of AI into educational frameworks holds immense potential to augment the efficiency, customization, and overall efficacy of educational practices. Such integration promises to cultivate a more dynamic and responsive educational environment, thereby facilitating enhanced learning experiences for students (Rosyadi et al., 2023). In adapting to the evolving landscape of education, it is imperative for learning media to embody characteristics that align with the principles of Society 5.0. These characteristics encompass adaptability, innovation, personalization, flexibility, and seamless integration with various cutting-edge technologies such as the Internet of Things (IoT), AI, and big data (Maulidia et al., 2023).

In the context of AI-integrated learning environments, various challenges in education arise, stemming from the inappropriate utilization of AI techniques, the evolving roles of teachers and students, and the emergence of social and ethical concerns (Zhai et al., 2021). Despite these challenges, research findings indicate that the majority of individuals strongly disagree with the notion that AI can replace teachers or lecturers, highlighting the enduring significance of human educators in the educational process (Pratama et al., 2023). Moreover, students demonstrate satisfaction with the incorporation of AI into learning activities, suggesting a positive reception and acceptance of AI-enhanced educational approaches (Rina, 2023). Furthermore, studies suggest that the integration of AI in elementary education yields favorable outcomes, including the enhancement of learning quality and the provision of an enriched learning experience (Mukti, 2023).

Additionally, AI-based blended learning models have been found to significantly impact students' science literacy, thereby contributing to the improvement of students' proficiency in scientific knowledge (Rahman et al., 2023). Furthermore, AI technologies offer the capability to automatically assess students' homework texts and deliver personalized recommendations based on their existing learning knowledge and domain expertise (Wang et al., 2020). These findings collectively underscore the evolving landscape of education, wherein AI plays a pivotal role in complementing traditional pedagogical methods while empowering both teachers and students in the learning process.

The integration of Artificial Intelligence (AI) into education stands as an imperative, necessitating the formulation of supportive policies, meticulous preparation, and an enhanced understanding of AI among educators. Furthermore, there exists a critical need for the development of adaptive curricula that can keep pace with technological advancements in various industries. Addressing these challenges proactively holds the promise of revolutionizing learning and teaching methodologies within the educational sphere, fostering an environment that is more inclusive, efficient, and of higher quality (Nuryadin & Marlina, 2023). As we explore practical implications for optimizing the positive impact of AI systems while mitigating negative consequences, it becomes evident that a balanced approach is necessary. Recommendations for maximizing benefits and minimizing drawbacks should be emphasized (Seo et al., 2021). Additionally, it is crucial to promote a comprehensive understanding of both the potential and limitations of AI. Educators and researchers play pivotal roles in harnessing AI's potential for educational purposes, while also recognizing and addressing the challenges it presents. By doing so, we can leverage

AI to enhance teaching practices effectively, capitalizing on its strengths while navigating its limitations (Rakuasa, 2023).

4. CONCLUSION

In conclusion, the research highlights the significance of utilizing visual stimuli in assessing Arabic speaking skills, emphasizing the integration of high-order thinking skills and the adoption of a comprehensive assessment approach. Through the incorporation of AI-generated images and the development of assessment protocols aligned with higher-order cognitive processes, educators can effectively evaluate students' proficiency in speaking Arabic. The assessment procedure follows standard protocols, integrating indicator matrices and grading rubrics to ensure objectivity and accuracy. Furthermore, the research underscores the multifaceted nature of speaking proficiency, encompassing pronunciation, grammar, vocabulary, fluency, and comprehension, all of which are crucial for language acquisition. It emphasizes the importance of oral tests over written ones in accurately assessing students' speaking abilities.

The integration of artificial intelligence (AI) into the production of images for speaking assessments presents a significant advancement in educational assessment methodologies. Utilizing platforms like the Bing AI Image Generator allows for the generation of tailored images to meet specific assessment needs, enhancing the effectiveness and efficiency of assessment practices. However, it is essential to recognize that the probability of successfully generating images that meet one's desires heavily relies on the manner of crafting prompts and the completeness of the prompts themselves. As AI users, individuals must engage in ample practice and exercise their creativity in formulating prompts as creatively and comprehensively as possible. Moreover, the provision of free access to such AI platforms by Bing enables widespread utilization by researchers and educators without financial barriers. As AI continues to evolve and permeate educational contexts, it becomes imperative to critically examine its implications, challenges, and ethical considerations. Nonetheless, AI also offers immense potential in revolutionizing digital media education and improving teaching and learning outcomes.

5. BIBLIOGRAPHY

- Adam, M. Z., Jundi, M., Ali, I., Laubaha, S. A., & Kadir, S. D. (2023). Studi Komparatif Kompetensi Dasar KMA 183 dan KMA 164 pada Mata Pelajaran Bahasa Arab Kelas VII MTs pada Kompetensi Ranah Kognitif. *Assuthur: Jurnal Pendidikan Bahasa Arab*, 2(1). <https://doi.org/10.58194/as.v2i1.262>
- Agarwal, M., Sharma, P., Goswami, A., Agarwal, M., Sharma, P., & Goswami, A. (2023). Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. *Cureus*, 15(6). <https://doi.org/10.7759/cureus.40977>
- Ainin, M. (2016). KESAHIHAN DALAM PENYUSUNAN TES BAHASA ARAB DI MADRASAH/SEKOLAH. *Prosiding Konferensi Nasional Bahasa Arab II, 1*, Article 2. <http://prosiding.arab-um.com/index.php/konasbara/article/view/75>
- Aiumtrakul, N., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Miao, J., Qureshi, F., & Cheungpasitporn, W. (2023). Navigating the Landscape of Personalized Medicine: The Relevance of ChatGPT, BingChat, and Bard AI in Nephrology Literature Searches. *Journal of Personalized Medicine*, 13(10), Article 10. <https://doi.org/10.3390/jpm13101457>
- Al-Ashwal, F. Y., Zawiah, M., Gharaibeh, L., Abu-Farha, R., & Bitar, A. N. (2023). Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools. *Drug, Healthcare and Patient Safety*, 15, 137–147. <https://doi.org/10.2147/DHPS.S425858>
- Ali, I., & Jundi, M. (2023). كتاب اللغة العربية للصف العاشر مدرسة عالية: التحليل بنظرية إعداد الكتاب التعليمي لناصر عبد الله الغالي. *JURNAL AL-IHDA: Media Ilmiah Bahasa Arab*, 11(2). <https://doi.org/10.58645/alihda.v11i2.418>
- Arikunto, S. (2009). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.
- Asrori, I., Thohir, M., & Ainin, M. (2019). *EVALUASI PEMBELAJARAN BAHASA ARAB* (A. F. Effendy, Ed.). MISYKAT.
- Baker, L. (2006). Observation: A Complex Research Method. *Library Trends*, 55(1), 171–189.
- Bernstein, J. (2006). Theory and method in automatic evaluation of spoken language proficiency. *The Journal of the Acoustical Society of America*, 120(5_Supplement), 3138. <https://doi.org/10.1121/1.4787747>
- Buaton, R., Fauzi, A., & Yel, M. (2022). New Paradigm E-Learning Model Based on Artificial Intelligence: New Paradigm E-Learning Model Based on Artificial Intelligence. *Journal of Artificial Intelligence and Engineering Applications*

- (JAIEA), 2(1), Article 1. <https://doi.org/10.59934/jaiea.v2i1.116>
- Cowie, N. (2009). Observation. In J. Heigham & R. A. Croker (Eds.), *Qualitative Research in Applied Linguistics: A Practical Introduction* (pp. 165–181). Palgrave Macmillan UK. https://doi.org/10.1057/9780230239517_8
- Dalle, M. (2019). تحليل الأخطاء النحوية والصرفية في تركيب الكلام: دراسة مقارنة بين برنامج الجدال العربي الكندي وبرنامج الجدال العربي منابر [Master, UIN Maulana Malik Ibrahim]. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=p--bHocAAAAJ&authuser=5&citation_for_view=p--bHocAAAAJ:9yKSN-GCB0IC
- Dinata, R. S., Musalwa, M., & Budiarti, M. (2021). Students' Perceptions about Evaluation of Speaking Skills (Mahârah Kalâm) Through Youtube Channel. *Lisaanuna Ta'lim Al-Lughah Al-Arabiyyah: Jurnal Pendidikan Bahasa Arab*, 4(2), 221–232. <https://doi.org/10.15548/lisaanuna.v4i2.3274>
- Fuentealba, C. (2011). The Role of Assessment in the Student Learning Process. *Journal of Veterinary Medical Education*, 38(2), 157–162. <https://doi.org/10.3138/jvme.38.2.157>
- Gipps, C. (1998). Student assessment and learning for a changing society. *Prospects*, 28(1), 31–44. <https://doi.org/10.1007/BF02737778>
- Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, 61(4), 5–14. <https://doi.org/10.1177/0008125619864925>
- Handayani, F., Nurhayati, N., & Kamila, A. (2022). Artificial intelligence as an educational media to improve adolescent reproductive health: Research and development studies. *Jurnal Keperawatan Padjadjaran*, 10(3), Article 3. <https://doi.org/10.24198/jkp.v10i3.2104>
- Hanief, R. (2022). Implementasi Model Penilaian Hots (Higher Order Thinking Skills) Pada Penilaian Empat Keterampilan Berbahasa Arab. *Ta'limi | Journal of Arabic Education and Arabic Studies*, 1(1), Article 1. <https://doi.org/10.53038/tlmi.v1i1.11>
- Hasibuan, R., & Jundi, M. (2023). استراتيجيات إدارية مبتكرة لتحقيق نجاح برنامج تعليم اللغة العربية في معهد دار العلوم الإسلامية في سيهاو. *Kitaba*, 1(3), Article 3. <https://doi.org/10.18860/kitaba.v1i3.24218>
- Hill, J. E., Harris, C., & Clegg, A. (2024). Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Research Synthesis Methods*, 15(2), 347–353. <https://doi.org/10.1002/jrsm.1689>
- Hosna, H., Firdaus, A. N. H., Mutmainnah, W., & Maulana, A. (2023). THE IMPLEMENTATION OF AUTHENTIC ASSESMENT IN

- RAISING THE STUDENT'S SPEAKING SKILL AT RELIGIOUS PROGRAMS (MANPK) SENIOR HIGH SCHOOL 01 JEMBER: *Ethical Lingua: Journal of Language Teaching and Literature*, 10(2). <https://doi.org/10.30605/25409190>. 637
- Ji, H., Han, I., & Ko, Y. (2022). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55, 48–63. <https://doi.org/10.1080/15391523.2022.2142873>
- Jundi, M. (2023a). Classical Test Theory in Analyzing Arabic Test Questions: A Descriptive Study on Item Analysis Research in Indonesia. *ATHLA: Journal of Arabic Teaching, Linguistic and Literature*, 4(2). <https://doi.org/10.22515/athla.v4i2.7747>
- Jundi, M. (2023b). تصميم برنامج أندرويد لتعزيز إتقان المفردات لدى طلاب المدرسة العالية. *Lahjah Arabiyah: Jurnal Bahasa Arab Dan Pendidikan Bahasa Arab*, 4(2), Article 2. <https://doi.org/10.35316/lahjah.v4i2.140-158>
- Jundi, M. (2023c). مقارنة تحليل بنود الأسئلة باستخدام نظرية الاختبار التقليدية ونظرية استجابة البنود لدرس اللغة العربية في المدرسة المتوسطة الإسلامية الحكومية باتو [Masters, Universitas Islam Negeri Maulana Malik Ibrahim]. <http://etheses.uin-malang.ac.id/60055/>
- Jundi, M., & Ali, M. (2023). Assessing The Classroom Learning in Arabic Matriculation Program: Tutor and Student Perspective. *Jurnal Al-Maqayis*, 10(2), 21–36.
- Jundi, M., & Hasibuan, R. (2023). Enhancing Arabic Language Proficiency among Students: A Case Study of Language Matriculation Strategies at Al-Hasyimiyah Darul Ulum Sipaho Islamic Boarding School. *Kilmatuna: Journal Of Arabic Education*, 3(2), Article 2. <https://doi.org/10.55352/pba.v3i2.619>
- Kementrian Pendidikan dan Kebudayaan. (n.d.). *Permendiknas No.16, Tahun 2007, tentang Standar Kualifikasi Akademik dan Kompetensi Guru*.
- Kholisoh, L. N. (2018). SUDAHKAH EVALUASI KEMAHIRAN BERBICARA BAHASA ARAB PADA TINGKAT DASAR DILAKUKAN? *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, Dan Budaya Arab*, 1(1), Article 1. <https://doi.org/10.17509/alsuniyat.v1i1.24200>
- Kshirsagar, P. R., Jagannadham, D. B. V., Alqahtani, H., Noorulhasan Naveed, Q., Islam, S., Thangamani, M., & Dejene, M. (2022). Human Intelligence Analysis through Perception of AI in Teaching and Learning. *Computational Intelligence and Neuroscience*, 2022, e9160727. <https://doi.org/10.1155/2022/9160727>
- Latief, M. A. (2016). Reliability of Language Skills Assessment Results. *Jurnal Ilmu Pendidikan*. <https://www.semanticscholar.org/p>

- aper/Reliability-of-Language-Skills-Assessment-Results-Latief/2b725100441a2500b3320498320bfc6cd27c34a0
- Mahmudi, I., & Nurbaha, A. (2023). Anwā'u al-Ikhtibārāt wa al-Mustawiyāt al-Mi'yāriyyah li Mahārah al-Kalām fī al-Lugah al-'Arabiyyah. *Aphorisme: Journal of Arabic Language, Literature, and Education*, 4(2), 112–127. <https://doi.org/10.37680/aphorisme.v4i2.3997>
- Mardapi, D. (2016). *Pengukuran Penilaian dan Evaluasi Pendidikan* (2nd ed.). Nuha Litera.
- Matrokhim, M. (2021). Students' Self-Assessment of Arabic Speaking Skill. *International Journal of Arabic Language Teaching*, 3(02), Article 02. <https://doi.org/10.32332/ijalt.v3i02.4208>
- Maulidia, L. N., Suparno, S., & Rosyidah, U. J. (2023). A Systematic Literature Review on Technology-Based Learning Media in ECE to Face Society 5.0 Era. *Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini*, 7(5), Article 5. <https://doi.org/10.31004/obsesi.v7i5.4997>
- Melati, S. E., Aini, N., & Wahyuni, A. F. (2023). The Uses of AI And Technology in Media and Material Development of Merdeka Belajar. *International Conference on Education*, 239–246.
- Miladya, J. (2015). EVALUASI DALAM PEMBELAJARAN BAHASA ARAB. *Prosiding Konferensi Nasional Bahasa Arab*, 1(1), Article 1. <http://prosiding.arab-um.com/index.php/konasbara/article/view/21>
- Miles, B. M., Huberman, A. M. and Saladana. (2014). *Qualitative Data Analysis A Methods Sourcebook. Edition 3* (p. 22). SAGE Publications.
- Mukti, F. D. (2023). TRANSFORMATION OF EDUCATION IN ELEMENTARY SCHOOLS: UTILIZATION OF ARTIFICIAL INTELLIGENCE-BASED LEARNING MEDIA IN THE DIGITAL ERA. *DIRASATUL IBTIDAIYAH*, 3(2), Article 2. <https://doi.org/10.24952/ibtidaiyah.v3i2.10200>
- Nuryadin, R., & Marlina, M. (2023). The Use of AI (Artificial Intelligence) in Education—Literature Review. *Indonesian Journal of Primary Education*, 7(2), Article 2. <https://doi.org/10.17509/ijpe.v7i2.64290>
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>
- Plough, I. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing*, 35(3), 325–329. <https://doi.org/10.1177/0265532218772322>
- Prananta, A. W., Susanto, N., Purwantoro, A., & Fuadah, N. (2023). ChatGPT Artificial Intelligence Integration

- in Science Learning Media: Systematic Literature Review. *Jurnal Penelitian Pendidikan IPA*, 9(7), Article 7. <https://doi.org/10.29303/jppipa.v9i7.4386>
- Pratama, M. P., Sampelolo, R., & Lura, H. (2023). REVOLUTIONIZING EDUCATION: HARNESSING THE POWER OF ARTIFICIAL INTELLIGENCE FOR PERSONALIZED LEARNING. *KLASIKAL: JOURNAL OF EDUCATION, LANGUAGE TEACHING AND SCIENCE*, 5(2), Article 2. <https://doi.org/10.52208/klasikal.v5i2.877>
- Rahman, A., Santosa, T. A., Ilwandri, I., Suharyat, Y., Aprilisia, S., & Suhaimi, S. (2023). The Effectiveness of AI Based Blended Learning on Student Scientific Literacy: Meta-analysis. *LITERACY: International Scientific Journals of Social, Education, Humanities*, 2(1), Article 1. <https://doi.org/10.56910/literacy.v2i1.542>
- Rakuasa, H. (2023). Integration of Artificial Intelligence in Geography Learning: Challenges and Opportunities. *Sinergi International Journal of Education*, 1(2), Article 2. <https://doi.org/10.61194/education.v1i2.71>
- Rina, N. (2023). Artificial Intelligence Use as Communication Media in Learning Activity for Senior High School. *Jurnal Ilmiah LISKI (Lingkar Studi Komunikasi)*, 9(2), 135–142. <https://doi.org/10.25124/liski.v9i2.6675>
- Rosser, H. L. (1979). Testing Oral Communicative Skills. *Foreign Language Annals*, 12(5), 371–373. <https://doi.org/10.1111/j.1944-9720.1979.tb00198.x>
- Rosyadi, M. I., Kustiawan, I., Tetehtio, E. O., & Joshua, Q. (2023). The Role of AI In Vocational Education: A Systematic Literature Review. *Journal of Vocational Education Studies*, 6(2), Article 2. <https://doi.org/10.12928/joves.v6i2.9032>
- Roviin, R. (2018). Evaluasi Pembelajaran Bahasa Arab (Kajian Tentang Instrumen Tes). *Arabia*, 10(1), Article 1. <https://doi.org/10.21043/arabia.v10i1.3085>
- Rukajat, A. (2018). *Pendekatan Penelitian Kualitatif (Qualitative Research Approach)*. Deepublish.
- Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18(1), 54. <https://doi.org/10.1186/s41239-021-00292-9>
- Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, 50(4), 1319–1330. <https://doi.org/10.1016/j.compedu.2006.12.005>
- Wang, W., & Liu, Z. (2021). Using Artificial Intelligence-Based Collaborative Teaching in Media

- Learning. *Frontiers in Psychology*, 12.
<https://www.frontiersin.org/journal/s/psychology/articles/10.3389/fpsyg.2021.713943>
- Wang, X., Gulenman, T., Pinkwart, N., De Witt, C., Gloerfeld, C., & Wrede, S. (2020). Automatic Assessment of Student Homework and Personalized Recommendation. *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, 150–154.
<https://doi.org/10.1109/ICALT49669.2020.00051>
- Wibisono, F., & Pramudita, D. A. (2023). *Smart geometry object detection: Flat shape recognition learning media based on artificial intelligence*. 020035.
<https://doi.org/10.1063/5.0151690>
- Widodo, J., & Qosim, M. N. (2021). Penilaian Kinerja Pada Pembelajaran Maharah Al-Kalam Level 1 Di Kursus Bahasa Arab Al-Arobiya Surakarta. *Uktub: Journal of Arabic Studies*, 1(2), Article 2.
<https://doi.org/10.32678/uktub.v1i2.5814>
- Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology*, 48(1), 1–5.
<https://doi.org/10.1080/17439884.2023.2167830>
- Yanhua, Z. (2020). The Application of Artificial Intelligence in Foreign Language Teaching. *2020 International Conference on Artificial Intelligence and Education (ICAIE)*, 40–42.
<https://doi.org/10.1109/ICAIE50891.2020.00017>
- Yasin, A., & Tarauni, H. (2023). Tashmim al-Ikhtibar li Maharah al-Kalam al-‘Arabi Wafqhan li al-Ithar al-Marja’i al-Urubi al-Musytarak li al-Lughat li al-Thullab al-Jami’ah fi Indunisiya. *Arabiyatuna: Jurnal Bahasa Arab*, 7(1 May), Article 1 May.
<https://doi.org/10.29240/jba.v7i1.6390>
- Yin, N. (2021). Research on the Impacts of Artificial Intelligence Technology on Language Teaching Innovation. *Frontiers in Educational Research*, 4(7).
<https://doi.org/10.25236/FER.2021.040706>
- Zaharo, Hangkiho, R., Jundi, M., Udin, R., & Hasibuan, R. (2024). Al-Ibtikār al-Hadīṣ fi al-Taḳyīm al-Takwīnī: Istifādātu min al-Ṣuwari al-Muwalladah bi al-Ḍakā’ al-Iṣṭinā’i fi Taḳyīm Maḥārah al-Kalām li al-Lughah al-Arabiyyah: الابتكار الحديث في التقييم التكويني: استفادة من الصور المولدة بالذكاء الاصطناعي في تقييم مهارة الكلام للغة العربية. *Al-Madrasah: Jurnal Pendidikan, Pembelajaran dan Kebudayaan*, 1(1), Article 1.
- Zaharo, Z., Chamidah, D., & Jundi, M. (2024). Can Word Puzzle-Based Formative Assessments Track Student Progress in Arabic Language Learning? *Kilmatuna: Journal Of Arabic Education*, 4(1), Article 1.
<https://doi.org/10.55352/pba.v4i1.871>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on

- artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 1–27. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021, e8812542. <https://doi.org/10.1155/2021/8812542>
- علام, ر. م. أ. (2005). *تقويم التعلم*. دار المسيرة للطباعة والنشر.
- عليما, م. م. (2022). بناء اختبار محكي المرجع في الرياضيات وفق النموذج اللوجستي ثلاثي المعلمة. *مجلة جامعة القدس المفتوحة للأبحاث والدراسات التربوية والنفسية*, 13(38), Article 38. <https://doi.org/10.33977/1182-013-038-004>
- عواضة, ه. (2010). *تقويم التعلم*. دار العلم للملايين.
- عودة, أ. (2005). *القياس والتقويم في العملية التدريسية*. شركة الأمل للطباعة والنشر والتوزيع.