Evaluating Visual Accuracy in AI-Generated Images of Malaysian-themed Icons

[1]Azahar Harun*, [2]Mohd Zaki Mohd Fadil, [3]Tengku Shahril Norzaimi Tengku Hariffadzillah

[1,2,3]Graphic Design Department, Faculty of Art and Design, Universiti Teknologi MARA Cawangan Melaka, Melaka, Malaysia

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.10100164

Received: 14 January 2026; Accepted: 19 January 2026; Published: 28 January 2026

ABSTRACT

Generative AI models are becoming widely accessible, enabling users across diverse backgrounds to create unprecedented artwork. However, this accessibility raises questions regarding the accuracy with which these models portray real-world subjects. Therefore, this paper examines three prominent generative AI models—Midjourney, DALL-E, and Stable Diffusion—to evaluate their efficacy in generating images of specific Malaysian-themed icons. Utilizing simple text prompts, the research phase was rigorously recorded and evaluated by an expert panel based on the Visual Appeal Rating Scale (VARS), encompassing eight criteria: Reliability, Consistency, Credibility, Professionalism, Aesthetics, Artistry, Harmony, and Balance. The results of the study indicate notable differences in model performance depending upon subject complexity. Midjourney emerged as the preeminent leader (Overall Mean: 3.25), exhibiting remarkable skill in culinary portrayal, attaining "Near Perfect" expert agreement on the aesthetics of the Nasi Lemak images. Stable Diffusion achieved a close second place (Overall Mean: 3.23), demonstrating proficiency in managing intricate structural geometry (Landmarks) and portraiture; yet, its elevated scores frequently coincided with "Slight" agreement, signifying considerable subjectivity in its technical performance. DALL-E was positioned third as a generalist model, yielding balanced albeit frequently contentious outcomes among specialists. A significant "Cultural Accuracy Gap" was identified across all models, wherein the representation of particular cultural icons (Politician) and intricate architecture (Landmark) was considerably more difficult than that of broad subjects (Food). DALL-E demonstrated significant inability in depicting the Malaysian politician, due to ethical concern. The study indicates that the existing generative AI models are specialized rather than universal; achieving high visual fidelity necessitates the deliberate selection of the model most appropriate for the specific aesthetic or structural requirements of the assignment.

Keywords: Creative arts, Digital images, Generative AI models, Inter-Rater Reliability, Malaysian-themed Icons

INTRODUCTION

A report by the World Bank in 2021 stated that the use of AI technology in Malaysia is increasing significantly in recent years. The escalating investments and interest in the technology from both the public and private sectors are a clear indicator of this. The Malaysian government recognizes the capacity of AI to drive economic growth and proactively promote the development and application of AI within the country. It is anticipated that this action will position Malaysia as one of key players in the global AI agenda.

In regard to the art and design sector, the use of generative AI models has significantly transformed and revolutionized traditional artistic practices. Among notable early generative AI models are MidJourney, Stable Diffusion, and DALL-E which are capable of generating images that align with user expectations for creativity and accuracy (Califano & Spence, 2024). Owing to these attributes, it is no wonder why generative AI models are increasingly being sought after by marketers and advertising agencies. Having said that, not all factors support the adoption of generative AI. For instance, the recent National Day advertisement displayed on an

electronic billboard in Desa Sri Hartamas, Kuala Lumpur, attracted viral attention on social media due to significant errors by the sponsor. A local news outlet reported that the generative AI employed by the advertising agency, WOW Media, "mistakenly" depicted the Petronas Twin Towers in Kuala Lumpur's city center with an additional third tower (The Sun Daily, 2024, November 10). Although regarded as a minor issue, the AI-generated Petronas "Triplet" Towers provoked considerable backlash, particularly from local netizens, who perceived it as unpatriotic, especially during National Day celebrations. Consequently, WOW Media was directed to retract the advertisement. This incident highlights the need for a critical evaluation of generative AI's role in the creative industries to ensure accuracy, credibility, and responsible usage.

Research Aims

This study has two aims. The first step is to investigate how generative AI models operate. The second step is to evaluate the visual accuracy of these AI-generated representations. Specifically, it seeks to determine the efficacy of generative AI models in producing digital images of Malaysian-themed icons and evaluate the extent to which these models accurately depict them.

LITERATURE REVIEW

Efforts to develop the concept of AI technology have been ongoing since the 1950s. In fact, it was first inspired from Hollywood science fiction films that featured intelligent robots (Anyoha, 2017). Classic films such as *Metropolis* (1927) and *The Wizard of Oz* (1939) are two examples that demonstrated the possibilities of artificial intelligence to harmoniously coexist in the human world.

The obsession and dream of smart machines were finally realized by a mathematics and computer scientist in 1950, named Alan Turing (Anyoha, 2017). In an article titled Computing Machinery and Intelligence, Turing developed a hypothesis that if the machine could access all forms of information, it would be impossible for it to solve various problems as humans. 5 years later an AI program called Logic Theorist was developed by Allen Newell, Cliff Shaw and Herbert Simon.

The Logic Theorist project has been funded by the Research and Development Corporation (RAND) and AI developed to mimic critical thinking human capabilities (Anyoha, 2017). For example, in 1997, a computer called Deep Blue developed by Intelligent Business Machine (IBM) successfully defeated the then chess champion Garry Kimovich Kasparov three times. IBM claimed that the success of Deep Blue has proven that true machines can really think on their own as well as set up strategies for overcoming human intelligence (Hankey, 2021).

Generative AI in creative arts

According to Newton and Dhole (2023), generative AI models are becoming more popular which enable artworks to be generated independently with the input of human artists. In contrast, Ramesh et al. (2021) contend that these models' capacity to produce unique and captivating visual content may disrupt the traditional creative process. According to Messer (2024), the generative AI models perform by mimicking specific painting techniques or creating entirely new compositions effortlessly, thus demonstrating an extensive array of creative potentials (Srinivasan, 2021). Presently, the generative AI models attract significant user interest such as Midjourney, Stable Diffusion, and DALL-E, owing to their capabilities to produce numerous variations of an idea within seconds (Smith et al., 2023).

Midjourney, Stable Diffusion & DALL-E

Midjourney, which was founded by David Holz is designed to produce drawings from narrative text prompts, a capability that may be very beneficial for developing contextual illustrations in diverse creative endeavors. The extensive utilization of this technology has ignited discussions around the originality and ownership of AI-generated images, since certain artists express apprehensions about their works being employed to train these generative AI models without permission (Nolan, 2023). DALL-E, created by OpenAI, is a notable model that transforms textual descriptions into visuals, akin to Midjourney, and has demonstrated potential uses in creative

sectors, such as advertising and education (Radford, et al., 2019). Stable Diffusion was created by Stability AI which is designed to provide accessible and transparent AI tools that democratize the creative capabilities of generative models. Unlike some previous models, Stable Diffusion's code and model weights are publicly available and can be run on most consumer hardware (Rombach et al., 2022).

Despite the advantages that generative AI offers, a fundamental question persists: should artists and designers normalize generative AI models as a standard practice in the art creation process? Some scholars argue that generative AI models include a condensed representation of millennia of human artistic endeavors, which holds significant relevance for art education (Dehouche & Dehouche, 2023). Others suggest that the involvement of human and generative AI models in art creation may affect its aesthetic and artistic worth. Although co-created work is regarded as more innovative, it lacks artistic authenticity, which exerts a prevailing impact (Messer, 2024). The findings indicate that artists' perceptions are adversely affected by the co-creation process, leading to diminished admiration for those who engage in co-creation, since they are viewed as less real.

METHODOLOGY

This study employed an exploratory research method to examine AI generated images by following the four main steps (as shown in Figure 1) including Text Prompt, Dataset, Generative AI Models, AI Artwork (Mazzone & Elgammal, 2019). Subsequently, these steps were modified to better align with the specific research questions and objectives of the study.

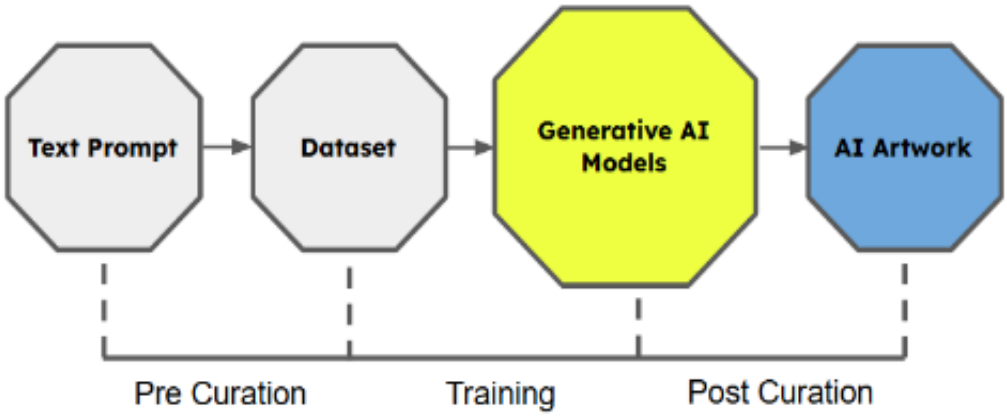


Figure 1. Four main steps of the generative AI process.

Malaysian-themed Icons

The selection of Malaysian-themed icons (Figure 2) was influenced by their cultural importance and popularity, together with the accessibility of pertinent visual data in generative AI training methods available online. According to the 2020 YouGov survey "Malaysia's Most Admired," The Honourable Tun Dr. Mahathir Mohamad was chosen for his prominence as a distinguished political figure in Malaysia, acknowledged both domestically and globally. His unique traits and continuous public presence render him an exemplary subject for assessing the effectiveness of generative AI models in representing notable individuals. The Petronas Twin Towers was selected for its significance as an iconic architectural landmark and a global symbol of Kuala Lumpur, frequently represented in media, tourism literature, and visual culture (Malaysian Tourism Promotion Board, n.d.). This makes them suitable for assessing the accuracy with which generative AI can reproduce complex structural and architectural elements. As reported by Michelin guide correspondent Yeoh (2024), Nasi Lemak is considered as Malaysia's popular dish and its deep association with the nation's cultural identity. Hence, its popularity and distinctive presentation provide a substantial assessment of generative AI models' capacity to comprehend and produce representations of traditional cuisine.

Collectively, these three Malaysian-themed icons category (Politician, Landmark, and Food) embody distinct yet readily identifiable facets of Malaysian culture, providing a comprehensive basis for assessing the precision, dependability, and cultural awareness of generative AI image outputs.

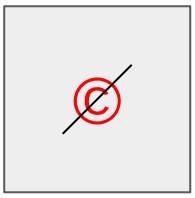
(i) (ii) (iii)

Figure 2. Malaysian-themed Icons from left to right: i. Politician: Tun Dr. Mahathir Mohamad; ii. Landmark: Petronas Twin Towers; iii. Food: Nasi Lemak

Exploration process

This study examined three generative AI models—Midjourney, DALL-E, and Stable Diffusion—to produce digital graphics of Malaysian-themed symbols based on a straightforward text prompt. “Hello, kindly create an image featuring a Malaysian icon: 1. Tun Dr. Mahathir Mohamad, 2. Petronas Towers, and 3. Nasi Lemak.” Nine images were produced from this method, as illustrated in Table 1, enabling the researcher to assess the efficacy of each generative AI model in producing digital representations.

Table 1. Malaysian themed icon generated by Midjourney, Stable Diffusion and DALL-E

No	Category	Original Source	Midjourney	Stable Diffusion	DALL-E
1	Politician				
2	Landmark				
3	Food				

Evaluation process

A Visual Appeal Rating Scale (VARS) was adapted and modified from the Aesthetic Scale (AS) developed by Lavie and Tractinsky (2004), which was originally designed to assess how users evaluate the visual look and feel of websites beyond usability, emphasizing aesthetics as an independent factor influencing user experience. For the purpose of this study, the scale was modified to suit the evaluation of AI-generated images of Malaysian-themed icons, with eight assessment criteria incorporated: Reliability, Consistency, Believability, Professionalism, Aesthetics, Artistic Quality, Harmony, and Balance (as shown in Table 2). Responses were measured using a five-point Likert scale questionnaire. This inductive approach helps define uncertain constructs

and guides item generation (Tay & Jebb, 2016). Six experts from the department of photography from Universiti Teknologi MARA Cawangan Melaka, Malaysia were chosen to evaluate the AI-generated images of Malaysian-themed icons in accordance with VARS. They received training on these criteria and were instructed in the utilization of a standardized evaluation form to guarantee uniformity in their assessments. Cohen's Kappa (κ) was used to measure inter-rater reliability (IRR), and provided percent agreement among evaluators (Krippendorff, 1970; Cohen, 1960). Three well known Malaysian-themed icons: Politician, Tun Dr. Mahathir Mohamad; Landmark, the Petronas Twin Towers; and Food, Nasi Lemak, were utilized as samples in the study (Figure 2.)

Table 2. Visual Appeal Criteria

No	Criteria	Questions
1	Reliability	How satisfied are you with the accuracy and reliability of the Malaysian icon's representation?
2	Consistent	How satisfied are you with the consistency of style and quality across all elements?
3	Believable	How satisfied are you with the realism and believability of the image?
4	Professional	How satisfied are you with the professional level of detail and execution displayed?
5	Aesthetic	How satisfied are you with the overall aesthetic appeal of the design and composition?
6	Artistic	How satisfied are you with the creativity and artistic originality of the work?
7	Harmony	How satisfied are you with the visual harmony between the colors, shapes, and proportions?
8	Balance	How satisfied are you with the visual balance and distribution of weight in the composition?

Percent Agreement and Inter-Rater Reliability (IRR)

Percent Agreement and Inter-Rater Reliability (IRR) are essential tools utilized to measure the degree of agreement among several evaluators. According to Stemler (2004) Percent Agreement is a frequentist metric that quantifies the direct ratio of instances where raters assign the same score, offering a clear and clearly interpretable index of absolute consensus. Cohen's Kappa was used to measure IRR in the image and assess the percent agreement among raters (Krippendorff, 1970; Cohen, 1960). The formula for calculating Kappa is as follow:

$$\text{kappa } (\kappa) = \frac{(Po - Pe)}{(1 - Pe)}$$

Note: Po= Observed Agreement; Pe = Expected Agreement

The interpretation of Kappa values and strength of agreement can be delineated as follows: scores ≤ 0.00 signify no agreement; values ranging from 0.01 to 0.20 indicate slight agreement; 0.21 to 0.40 reflect fair agreement; 0.41 to 0.60 represent moderate agreement; 0.61 to 0.80 denote substantial agreement; and 0.81 to 1.00 correspond to near-perfect agreement (McHugh, 2012). This interpretative scale facilitates the assessment of rating consistency and dependability, hence strengthening the evaluation process.

ANALYSIS & RESULT

MidJourney

Table 3. Politician: Tun Dr Mahathir Mohamad

Criteria	Expert Scores (E1–E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	2, 2, 2, 2, 1, 3	2	66.70%	2	0.667	Substantial
Consistent	2, 2, 2, 2, 2, 4	2	83.30%	2.33	0.667	Substantial
Believable	2, 2, 2, 2, 2, 3	2	83.30%	2.17	0.4	Moderate / Fair
Professional	2, 2, 3, 2, 2, 4	2	66.70%	2.83	0.4	Moderate / Fair
Aesthetic	2, 3, 4, 3, 4, 4	4	50.00%	3	0.267	Fair
Artistic	2, 3, 2, 2, 3, 4	2	50.00%	2.67	0.267	Fair
Harmony	2, 4, 4, 2, 5, 4	4	50.00%	3.67	0.133	Slight / Poor
Balance	2, 4, 3, 2, 5, 4	4	33.30%	3.33	0.133	Slight / Poor

Table 3 shows the dataset of the Midjourney generated image of Politician: Tun Dr Mahathir Mohamad. The evaluation reveals a significant divergence in expert opinion between technical accuracy and visual appeal: while the experts largely agree on the output's limitations (Mode score of 2 for most technical criteria), they disagree on its artistic merit. For categories like Reliability and Consistency, the agreement strength is Substantial (κ = 0.667), indicating a high degree of consensus that the image failed to meet high standards of likeness. However, for Harmony and Balance, the agreement strength drops to Slight Poor (κ = 0.133), with a very low agreement rate of 33.30% for the latter. While the majority of experts converged on a low score for "Believability," their opinions on "Aesthetic" and "Harmony" were far more scattered, with scores ranging from 2 to 5. This suggests that while Midjourney failed to produce a technically accurate portrait, experts were divided on whether the resulting image possessed incidental artistic value, making it a "technically poor but visually divisive" output.

Table 4. Landmark: Petronas Twin Towers

Criteria	Expert Scores (E1–E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	2, 1, 4, 2, 4, 3	2 or 4	33.33%	2.67	-0.083	Poor
Consistent	2, 1, 4, 2, 4, 4	4	50.00%	2.83	0.083	Slight
Believable	2, 1, 2, 2, 3, 4	2	50.00%	2.33	0	Slight / Neutral
Professional	2, 2, 2, 2, 3, 4	2	66.67%	2.5	0.25	Fair
Aesthetic	2, 3, 2, 2, 4, 4	2	50.00%	2.83	0.083	Slight

Artistic	2, 3, 2, 2, 4, 4	2	50.00%	2.83	0.083	Slight
Harmony	2, 4, 4, 2, 5, 4	4	50.00%	3.5	0.083	Slight
Balance	2, 3, 4, 2, 5, 4	2	33.33%	3.33	-0.083	Poor

Table 4 shows the dataset of the Midjourney generated image of Landmark: Petronas Twin Towers. The evaluation presents a significant challenge in consensus, as the experts largely disagree on the quality and execution of the landmark. Unlike other subjects that leaned toward a clear success or failure, this output yielded a low Mode score of 2 for the majority of categories, including Believability, Aesthetics, and Artistic merit.

The statistical strength of this evaluation is notably weak, with agreement levels categorized primarily as "Slight" or "Poor." For critical categories like Reliability and Balance, the Kappa values are negative (κ = -0.083), indicating that the experts agreed even less than what would be expected by random chance. While Professionalism reached a Fair agreement (κ = 0.25), the overall data suggests the image was highly inconsistent. Experts were deeply divided; for instance, in the Harmony category, scores ranged from a low of 2 to a maximum of 5. Ultimately, the Midjourney output for this landmark can be described as an "unsuccessful and polarizing" representation, failing to achieve a reliable or professional standard in the eyes of the expert panel.

Table 5. Food: Nasi Lemak

Category	**Expert Scores (E1–E6)**	**Mode (Score)**	**Percent Agreement**	**Mean Score**	**IRR Kappa Value (κ)**	**Agreement Strength**
Aesthetic	4, 4, 4, 4, 4, 4	4.00	100.00%	4.00	1.00	Near Perfect
Reliability	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Consistent	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Believable	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Professional	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Artistic	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Harmony	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate
Balance	4, 4, 4, 4, 5, 4	4.00	83.33%	4.17	0.58	Moderate

Table 5 shows the dataset of the Midjourney generated image of Food: Nasi Lemak. In contrast to the previous landmark analysis, this evaluation demonstrates an exceptionally high level of expert consensus and a strong positive reception. The imagery achieved a unanimous Mode score of 4 across every category, with Mean scores slightly exceeding 4.00 due to a "Very Satisfied" rating (5) from one expert (R5). The statistical robustness of this dataset is significant. In the Aesthetic category, the panel reached a 100% Percent Agreement (κ = 1.00), classified as "Near Perfect." All other categories, including Reliability, Professionalism, and Harmony, maintained a high agreement rate of 83.33% and a Moderate agreement strength (κ = 0.58). This indicates that Midjourney was highly successful in producing a representation that experts found consistently appealing and

technically sound. The presence of only one slight variation in scoring suggests that the output is a "unified success," possessing clear, objective qualities that were easily recognized and valued by the entire expert panel.

Stable Diffusion

Table 6. Politician: Tun Dr Mahathir Mohamad

Criteria	Expert Scores (E1–E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	4, 3, 2, 4, 5, 4	4.00	50.00%	3.67	0.00	Slight
Consistent	4, 3, 2, 4, 5, 4	4.00	50.00%	3.67	0.00	Slight
Believable	4, 3, 2, 4, 5, 4	4.00	50.00%	3.67	0.00	Slight
Professional	4, 3, 2, 4, 5, 4	4.00	50.00%	3.67	0.00	Slight
Aesthetic	4, 3, 3, 4, 5, 4	4.00	50.00%	3.83	0.08	Slight
Artistic	4, 4, 2, 4, 5, 4	4.00	66.67%	3.83	0.25	Fair
Harmony	4, 4, 3, 4, 5, 4	4.00	66.67%	4.00	0.25	Fair
Balance	4, 3, 3, 4, 5, 4	4.00	50.00%	3.83	0.08	Slight

Table 6 shows the dataset of the Stable Diffusion generated image of Politician: Tun Dr Mahathir Mohamad. The evaluation presents a generally positive reception with a consistent Mode score of 4.00 across all categories, indicating that the majority of experts found the depiction satisfactory. The Mean scores, ranging from 3.67 to 4.00, further support a favorable lean in the overall assessment. However, the statistical reliability of these scores is notably low. For the foundational categories of Reliability, Consistency, Believability, and Professionalism, the agreement strength is categorized as "Slight" ($\kappa = 0.00$). This indicates that while half the experts agreed on a score of 4, the overall distribution of scores (ranging from a low of 2 to a high of 5) makes the agreement statistically no better than random chance. The categories of Artistic merit and Harmony performed slightly better, achieving a "Fair" strength of agreement ($\kappa = 0.25$) with a 66.67% agreement rate. In conclusion, while Stable Diffusion succeeded in creating a visually acceptable image, the high variance in expert scores suggests that the output is "subjectively pleasing but technically inconsistent," as experts could not reach a robust consensus on its objective quality.

Table 7. Landmark: Petronas Twin Towers

Criteria	Expert Scores (E1–E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	4, 3, 3, 4, 5, 4	4.0	50.00%	3.83	0.08	Slight
Consistent	4, 3, 4, 4, 5, 4	4.0	66.67%	4.00	0.25	Fair
Believable	4, 3, 4, 4, 5, 4	4.0	66.67%	4.00	0.25	Fair
Professional	4, 4, 4, 4, 5, 4	4.0	83.33%	4.17	0.58	Moderate

Aesthetic	4, 4, 4, 4, 5, 4	4.0	83.33%	4.17	0.58	Moderate
Artistic	4, 4, 4, 4, 5, 4	4.0	83.33%	4.17	0.58	Moderate
Harmony	4, 4, 4, 4, 5, 4	4.0	83.33%	4.17	0.58	Moderate
Balance	4, 4, 3, 4, 5, 4	4.0	66.67%	4.00	0.25	Fair

Table 7 shows the dataset of the Stable Diffusion generated image of Landmark: Petronas Twin Towers. This result demonstrates a highly successful and cohesive reception among the expert panel, with a unanimous Mode score of 4 across all eight criteria. The Mean scores are notably high, ranging from 3.83 to 4.17, which indicates that the model consistently met or exceeded the experts' expectations for a satisfactory landmark representation. The statistical strength of this evaluation is particularly robust in the categories of Professionalism, Aesthetics, Artistic merit, and Harmony. These areas achieved an 83.33% level of agreement and a Moderate agreement strength (κ = 0.58), suggesting that the visual and technical quality of the generated towers was clear and objective to nearly the entire panel. While categories such as Consistency, Believability, and Balance showed a Fair strength of agreement (κ = 0.25), the overall data reveals a strong alignment. Only the Reliability category showed a Slight agreement strength (κ = 0.083), primarily due to a wider spread of Expert scores (3 to 5). In summary, the Stable Diffusion output for this landmark is a "highly credible success," characterized by high Mean scores and a significant degree of expert agreement regarding its professional and aesthetic quality.

Table 8. Food: Nasi Lemak

Criteria	**Expert Scores (E1–E6)**	**Mode (Score)**	**Percent Agreement**	**Mean Score**	**IRR Kappa Value (κ)**	**Agreement Strength**
Reliability	1, 2, 2, 1, 1, 4	1.00	50.00%	1.83	0.08	Slight
Consistent	1, 2, 2, 1, 1, 4	1.00	50.00%	1.83	0.08	Slight
Believable	1, 1, 2, 1, 1, 4	1.00	66.67%	1.67	0.25	Fair
Professional	1, 2, 2, 1, 1, 4	1.00	50.00%	1.83	0.08	Slight
Aesthetic	2, 2, 2, 1, 1, 4	2.00	50.00%	2.00	0.08	Slight
Artistic	1, 2, 2, 1, 1, 4	1.00	50.00%	1.83	0.08	Slight
Harmony	1, 2, 2, 1, 1, 4	1.00	50.00%	1.83	0.08	Slight
Balance	2, 2, 2, 1, 1, 4	2.00	50.00%	2.00	0.08	Slight

Table 8 shows the dataset of the Stable Diffusion generated image of Food: Nasi Lemak. The evaluation reveals a predominantly negative reception from the expert panel, representing a significant failure in the model's ability to depict this subject accurately. The results are characterized by very low Mean scores, ranging from 1.67 to 2.00, and a Mode score of 1 for most criteria, indicating "Strong Dissatisfaction" among the majority of experts. The statistical reliability of this assessment is categorized almost entirely as "Slight." For seven out of eight categories, the Kappa value is = 0.08. While the Believable category reached a "Fair" strength of agreement (κ = 0.25) due to a 66.67% agreement rate on the lowest score, the overall consensus remains weak. The data is significantly skewed by a single outlier, Expert 6 (E6), who consistently provided an Expert score of 4

(Satisfied), while the rest of the panel provided Expert scores of 1 or 2. This stark contrast suggests that while one expert found the representation acceptable, the overwhelming majority viewed it as an "unsuccessful and unconvincing" output, failing both the technical and aesthetic requirements of the prompt.

DALL-E

Table 9. Politician: Tun Dr Mahathir Mohamad

Criteria	Expert Scores (E1-E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Consistent	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Believable	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Professional	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Aesthetic	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Artistic	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Harmony	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect
Balance	1, 1, 1, 1, 1, 1	1.00	100.00%	1.00	1.00	Near Perfect

The dataset in Table 9 shows the analysis of the DALL-E generated image of Politician: Tun Mahathir Mohamad. Across every criterion—Reliability, Consistency, Believability, Professionalism, Aesthetic, Artistic, Harmony, and Balance—every single expert (E1 through E6) assigned the lowest possible score of 1 (Expert Score: 1 (Very Dissatisfied); Agreement Strength: Near Perfect (κ = 1.0). Statistically, this result implies that the generated output for this subject failed so clearly that no expert found even minor redeeming qualities to warrant a higher score. The reason for this is that DALL-E was unable to generate the image owing to ethical concerns.

Table 10. Landmark: Petronas Twin Towers

Criteria	Expert Scores (E1–E6)	Mode (Score)	Percent Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	3, 4, 2, 5, 4, 4	4.00	50.00%	3.67	0.00	Slight
Consistent	4, 4, 2, 5, 4, 4	4.00	66.67%	3.83	0.25	Fair
Believable	3, 4, 2, 5, 4, 4	4.00	50.00%	3.67	0.00	Slight
Professional	4, 4, 2, 5, 4, 4	4.00	66.67%	3.83	0.25	Fair
Aesthetic	4, 4, 2, 5, 4, 4	4.00	66.67%	3.83	0.25	Fair
Artistic	3, 4, 2, 5, 4, 4	4.00	50.00%	3.67	0.00	Slight
Harmony	3, 4, 2, 5, 4, 4	4.00	50.00%	3.67	0.00	Slight
Balance	3, 4, 2, 5, 4, 4	4.00	50.00%	3.67	0.00	Slight

Table 10 shows the dataset of the DALL-E generated image of Landmark: Petronas Twin Towers. The evaluation presents a paradox: despite general positive sentiment (Mode score of 4), the experts strongly disagree on the details. For categories like Reliability, Believability, Artistic, Harmony, and Balance, the agreement strength is statistically no better than random chance (κ = 0 (Slight) and κ = 0.25 (Fair). While the majority of experts "liked" the result, they likely liked it for completely different reasons. The visual representation was polarizing; some found it highly artistic (5), while others found it lacking (2).

Table 11. Food: Nasi Lemak

Criteria	Expert Scores (E1–E6)	Mode (Score)	% Agreement	Mean Score	IRR Kappa Value (κ)	Agreement Strength
Reliability	3, 4, 4, 4, 2, 4	4.00	66.67%	3.50	0.25	Fair
Consistent	3, 4, 4, 4, 2, 4	4.00	66.67%	3.50	0.25	Fair
Believable	3, 4, 4, 4, 2, 4	4.00	66.67%	3.50	0.25	Fair
Professional	4, 4, 4, 4, 2, 4	4.00	83.33%	3.67	0.58	Moderate
Aesthetic	3, 4, 4, 4, 2, 4	4.00	66.67%	3.50	0.25	Fair
Artistic	4, 4, 4, 4, 2, 4	4.00	83.33%	3.67	0.58	Moderate
Harmony	4, 4, 4, 4, 2, 4	4.00	83.33%	3.67	0.58	Moderate
Balance	3, 4, 4, 4, 2, 4	4.00	66.67%	3.50	0.25	Fair

Table 11 shows the dataset of the DALL-E generated image of Food: Nasi Lemak. The evaluation presents a more stable consensus compared to the Politician and the landmark subject: while the general sentiment remains positive (Mode score of 4), the experts show a higher level of alignment. For categories like Professional, Artistic, and Harmony, the agreement strength is statistically significant, reaching a Moderate agreement rate of 83.33% (κ = 0.583). Other categories like Reliability, Believability, and Aesthetic produced a Fair agreement (κ = 0.25). While the majority of experts (5 out of 6) were in perfect alignment for the top-performing categories, the overall agreement was slightly tempered by a single outlier (E5) who consistently rated the dish 2 (Dissatisfied). Unlike the polarizing nature of the landmark imagery, the Nasi Lemak representation is a "robust success," characterized by strong group alignment and a clear, high-quality expert verdict.

FINDINGS & DISCUSSION

The assessment of generative AI models reveals a complex scenario where technical capability and cultural accuracy often diverge. While all three models demonstrate the potential to produce high-quality visual content, an analysis of Percent Agreement and Inter-Rater Reliability (IRR) proves that their efficacy is significantly dictated by the subject matter. These findings suggest that the generative AI model is currently divided into specialized domains rather than a single, dominant leader.

Technical Reliability and Expert Agreement

The statistical data reveals a sharp contrast in how "believable" or "standardized" the outputs are across models. Midjourney emerged as the most consistent performer, setting a benchmark for aesthetic quality. Its representation of Nasi Lemak, for instance, reached a "Near Perfect" agreement among experts. This level of consensus implies that Midjourney’s training data for culinary subjects is highly aligned with professional standards, making it the most dependable model for gastronomic imagery.

In contrast, Stable Diffusion presents a paradox of subjectivity. While it frequently achieved high scores (Mode 4) for both politicians and landmarks, the expert agreement was often categorized as "Slight." This discrepancy indicates that while the model creates visually appealing art, it frequently lacks the technical precision required to satisfy a diverse panel of experts.

DALL-E occupied a more unpredictable position, exhibiting the highest degree of disagreement. Its performance was characterized by extreme division, manifesting as either a total, unanimous failure in portraying specific public figures or a "controversial success" in architectural tasks. This instability suggests that DALL-E's outputs often leave experts deeply divided on whether the result is an intentional artistic interpretation or a fundamental error.

Navigating the "Cultural Gap"

A recurring theme throughout the study is the "Cultural Gap," where models struggled significantly more with specific cultural icons and landmarks than with general objects like food. This gap is most evident in three key areas:

Cultural Likeness: Portraying specific figures like Tun Dr. Mahathir Mohamad proved to be the most challenging undertakings. Stable Diffusion outperformed DALL-E in this regard, though the "Slight" agreement suggests it captured a recognizable "ambiance" rather than a precise, believable likeness. DALL-E's failure in this area was absolute, underscoring the limitations of general-purpose models in maintaining cultural fidelity. In contrast, DALL-E's inability to generate a likeness was not due to technical limitations, but rather a deliberate design decision. OpenAI has established stringent safety measures and ethical safeguards, specifically a "multi-tiered safety system", to restrict the creation of images of famous persons by name, hence reducing the hazards of misinformation and deepfakes. Thus, DALL-E's incapability to depict the politician signifies a delicate compliance with its Content Policy rather than a shortcoming in its foundational data integrity. This underscores a significant divide within the industry: certain generative AI models facilitate the examination of cultural resemblance, whilst others emphasize ethical protections, substantially transforming their approach to activities related to "cultural fidelity."

Structural Complexity: Stable Diffusion demonstrated a superior understanding of structural geometry, particularly with the Petronas Twin Towers. It attained a "Moderate" consensus in professionalism and aesthetics, suggesting it is better equipped for architectural visualization. DALL-E's interpretation of the same structure was regarded as "unsuccessful" and "poorly balanced," failing to replicate the landmark's iconic symmetry.

Gastronomic Representation: While Midjourney excelled in food photography, Stable Diffusion failed significantly in the same category. This suggests that Stable Diffusion's current model weights may struggle with the specific textural and compositional elements—such as the unique consistency of rice and side dishes—required for authentic Malaysian cuisine.

Table 12. Summary of Analysis

Rank	Model	Politician (Mean)	Landmark (Mean)	Food (Mean)	Overall Mean	Agreement Strength (Typical)
1	Midjourney	2.75	2.85	4.15	3.25	Moderate to Near Perfect
2	Stable Diffusion	3.77	4.06	1.85	3.23	Slight to Moderate
3	DALL-E	2.75	2.86	3.56	3.06	Poor to Substantial

CONCLUSION

The expert evaluation of AI-generated images of Malaysian-themed icons from Midjourney, Stable Diffusion, and DALL-E reveals a varied agreement where technical proficiency and cultural fidelity frequently disagree. Although all three models exhibit the capacity to generate high-quality visuals, this study establishes that their efficacy is largely influenced by the subject matter. Through percent agreement and inter-rater reliability analysis, the study determined that the generative AI models are currently segmented into specialized domains rather than being governed by a dominant model.

Midjourney stands as the aesthetic leader of this category, achieving the highest rank with an Overall Mean of 3.25. The performance in the "Food" category, particularly with the representation of Nasi Lemak, was outstanding. With a mean of 4.15 and a "Near Perfect" agreement strength (κ = 1.00), Midjourney has achieved a level of aesthetic consistency that is nearly indistinguishable from professional food photography. Nonetheless, this achievement is somewhat mitigated by its performance in intricate structural and portrait tasks, where it exhibited a greater level of rater subjectivity.

Stable Diffusion ranks a close second, with an Overall Mean of 3.23. This model had exceptional structural and spatial integrity, leading the "Landmark" category with a mean score of 4.06. It surpassed its competitors in encapsulating the "essence" of human pictures, as evidenced by the assessment of Tun Dr Mahathir Mohamad. A significant paradox was noted: whereas Stable Diffusion had excellent scores, its agreement strength was frequently classified as "Slight." This indicates that although the model produces visually "pleasing" images, they often lack the technical accuracy necessary for a cohesive expert agreement, rendering the outcomes aesthetically gratifying yet technically disagreeing.

DALL-E, with an Overall Mean of 3.06, ranks third and functions as an effective generalist. It sustained a consistent baseline across categories; yet, its performance was frequently characterized as "polarizing." For example, in the "Landmark" category, it produced a "controversial success," with experts sharply divided over its artistic worth and structural accuracy. The inability to generate the "Politician" category, despite "Substantial" expert consensus, underscores the persistent "Cultural Gap" across broader models. The selection of a generative model must be deliberate. For projects necessitating superior industrial aesthetics and gastronomic authenticity, Midjourney is the preeminent authority. Stable Diffusion provides the highest capabilities for architectural visualization and conceptual portraiture that emphasize structural depth. DALL-E is a versatile choice for general prototyping that necessitates a balance of diverse subjects.1 As these models advance, developers will face the difficulty of reconciling "subjective beauty" with "objective accuracy," especially with regional cultural symbols and intricate architectural features.

ACKNOWLEDGEMENT

The author wishes to acknowledge the support of the Institut Seni Kreatif Nusantara (INSAN), the Faculty of Art and Design, University Teknologi MARA Cawangan Melaka, as well as the contributions of co-authors and experts from the photography department who assisted in analyzing the efficacy of the generative AI technologies under study.

REFERENCES

1. Anyoha, R. (2017, August 28). The history of artificial intelligence. Science in the News. Harvard University. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence
2. Califano, G., & Spence, C. (2024). Assessing the visual appeal of real/AI-generated food images. Food Quality and Preference, 116, 105149. https://doi.org/10.1016/j.foodqual.2024.105149
3. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104
4. Dehouche, N., & Dehouche, K. (2023). What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon, 9(6), e16757. https://doi.org/10.1016/j.heliyon.2023.e16757

5. Hankey, A. (2021). Kasparov versus Deep Blue: An illustration of the Lucas–Gödelian argument. Cosmos and History: The Journal of Natural and Social Philosophy, 17(3), 60–67. https://www.cosmosandhistory.org/index.php/journal/article/view/989
6. Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. Sociological Methodology, 2, 139–150. https://doi.org/10.2307/270769
7. Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. Int. J. Hum. Comput. Stud., 60, 269-298.
8. Mazzone, M., & Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. Arts, 8(1), 26. https://doi.org/10.3390/arts8010026
9. Malayian Tourism Promotion Board (n.d). Famous Architectural Landmarks In Malaysia. https://www.malaysia.travel/explore/petronas-twin-tower#:~:text=The%20Petronas%20Twin%20Towers%20are%20a%20famous,Park**%20*%20**Aquaria%20KLCC**%20*%20**Kinokuniya%20Bookstore**
10. Messer, U. (2024). Co-creating art with generative artificial intelligence: Implications for artworks and artists. Computers in Human Behavior: Artificial Humans, 2(1), 100056. https://doi.org/10.1016/j.chbah.2023.100056
11. McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochem Med (Zagreb), 22(3), 276-282. PMID: 23092060; PMCID: PMC3900052.
12. Newton, A., & Dhole, K. (2023). Is AI art another industrial revolution in the making? arXiv. https://doi.org/10.48550/arxiv.2301.05133
13. Nolan, B. (2023, January 15). This man used AI to write and illustrate a children's book in one weekend. He wasn't prepared for the backlash. Business Insider. https://www.businessinsider.com/chatgpt-midjourney-ai-write-illustrate-childrens-book-one-weekend-alice-2023-1
14. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv. https://arxiv.org/abs/2102.12092
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684–10695). IEEE. https://doi.org/10.1109/CVPR52688.2022.01042
17. Srinivasan, R. (2021). Quantifying confounding bias in generative art: A case study. arXiv. https://doi.org/10.48550/arxiv.2102.11957
18. Smith, A., Schroeder, H., Epstein, Z., Cook, M., Colton, S., & Lippman, A. (2023). Trash to treasure: Using text-to-image models to inform the design of physical artefacts. arXiv. https://doi.org/10.48550/arXiv.2302.00561
19. Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(1), 4.
20. Tay, L., & Jebb, A. T. (2016). Scale development. In S. Rogelberg (Ed.), The SAGE encyclopedia of industrial and organizational psychology (2nd ed., Vol. 4, pp. 1365–1370). Sage.
21. The Sun Daily. (2024, November 10). Netizens outraged by Brickfields PDRM's AI Merdeka billboard depicting 3 KLCC towers. The Sun Daily. https://thesun.my/style-life/going-viral/netizens-outraged-by-brickfields-pdrm-s-ai-merdeka-billboard-depicting-3-klcc-towers-AI12837403
22. YouGov. (2020). Malaysia's most admired. https://yougov.com/articles/32284-malaysias-most-admired
23. Yeoh, P. (2024, September 5). Iconic dishes: Nasi lemak, the quintessential Malay breakfast. MICHELIN Guide. https://guide.michelin.com/my/en/article/features/what-is-nasi-lemak