



# Can Image Generative Models be Considered Experts?

Dang Ba Hai Quang <sup>1</sup>[0009-0009-0183-1651], Ginel Dorleon<sup>2</sup>[0000-0003-2343-4445], and Andrew Colarik<sup>3</sup>[0000-0001-5932-8201]

<sup>1,2,3</sup>RMIT University, Ho Chi Minh, Vietnam

 s3676330@student.rmit.edu.au

<sup>2,3</sup> `firstname.lastname@rmit.edu.vn`

**Abstract.** This paper addresses foundational challenges in evaluating Generative Artificial Intelligence (GAI), focusing on the transition from expertise evaluation to intelligence evaluation. It critiques both quantitative and qualitative metrics for GAI, highlighting limitations in human-algorithm interaction environments. The study examines knowledge representation in neural network architectures and the processes of filtering versus tokenisation for image processing, emphasising inconsistencies and lack of standardisation in test design. Based on this finding the paper proposes a framework for future research to further explore the research questions.

**Keywords:** Generative Artificial Intelligence · Image · AI · Intelligence Evaluation · Evaluation Framework

## 1 Introduction

The rapid advancement of Generative Artificial Intelligence (GAI) [3] can greatly contribute to the rapid introduction of novel text prediction models such as Generative Pre-train Transformer (GPT), Retrieval Augmented Generation (RAG), Bidirectional Encoder Representation from Transformer (BERT) and image generation architectures such as Generative Adversarial Network (GAN), Variation Auto Encoder (VAE) [19], Transformer [34] and the recent introduction of mixture models such as CLIPs [9]. These architectures demonstrate an incredible ability to replicate human responses and generate seemingly novel ideas and content - sound, image, and text - that showcase a new form of machine intelligence not available in traditional Machine Learning (ML) models. Propelled in part by media attention and through the accumulation of Artificial Intelligence (AI) research breakthroughs, these models are now jumping out of the traditional lab research environment into the hands of the layperson in the form of general chatbots and commercial applications [3, 21]. One such application is the use of GAI as a form of expert-based system in various domains such as helping with medical diagnosis [18], creating art [32] and provide recommendation for business [26].

However, alongside this application, there is the ongoing question of whether or not these systems possess an "expert" level of knowledge. Traditionally, computers have been an important tool in aiding humans by leveraging their computational ability to either seek to enhance or replicate tasks that were usually only reserved for expert [3]. However different from traditional ML where the algorithm is bound by some explainable mathematical function, current GenAI models learn to generate new information bases in a complex multilayer hidden network of artificial neurons. This inherently creates a black-box phenomenon where the output reasoning is not clearly understood easily by domain experts, developers and researchers [3]. In complex and high-risk environments such as medical, aerospace, or precision manufacturing this unreliability and lack of transparency is simply unacceptable. Thus, many researchers are trying to overcome this limitation by implementing hard rule-based limitations, physical scoring based on traditional programming, or adding humans to the decision-making and evaluation loop [28]. However, this area of research remains in its early stage because the field of GAI is often empirically based.

Currently, most research on knowledge embedding has focused primarily on text generation in large language models (LLMs), with comparatively little research on knowledge embedding in image generation [3]. Likely this is thanks to the massive success of ChatGPT, Stable Diffusion which pushes the public and research direction into more creative content generation. However, there is a gap in theoretical research on the generation of conditioning technical images such as engineering drawings, medical imaging, and more technical design [39]. This is particularly interesting considering there is a lot of research on the application of deep neural networks in these fields for high-precision tasks which often can contribute more value to the development of many industries.

## 2 Problem Definition

Goldman [10] defines expertise in humans as the ability to help others with the focus on laypersons to solve a variety of problems in a domain that they would not be able to solve on themselves. This can be either by directly performing the tasks or through indirect means such as providing training or advice. Thus, following this definition, we ask the question:

*Question 1.* Can generative image models contain domain-specific expertise by producing images with complex semantic information suitable for high-constraint fields?

In simpler terms, can the model generate a group of images that are both visually and functionally close to expert-generated images, which meet a performance threshold and provide explainability relevant to the domain where high expertise is required? We are interested in solving this problem as it would imply the ability to adopt these tools in high-constrain, precision, and low error-tolerance domains such as medical, engineering, aerospace and many precision manufacturing domains. While we believe it is unlikely to fully replace human

experts, the additional computational power would allow for more productive human-machine collaboration by leveraging these model capabilities to combine and generate novel ideas from collective knowledge.

### 3 Foundational Challenges

To better understand the existing similar research as well as the problem space, the paper conducts a non-systematic literature review on Deep Generative Neural Networks (DGNN) [3] with a focus on image generation. We found some foundational challenges that cause the problem to persist in the field.

#### 3.1 The Shift from Expertise Evaluation to Intelligence Evaluation

Traditionally expertise is evaluated in humans as the ability to rapidly recognise patterns from one domain and translate them into practice [7, 10]. This is a way for experts to outperform laypersons in their fields and allow for the advancement of their respective fields. Thus, it is often the reason why AI researchers are deeply interested in replicating these traits [27, 3]. However, early systems that use conditional programming are fixed and often inflexible to the new changes as the domain evolved [27]. Algorithm-based systems follow suit with an improvement in the ability to learn from existing patterns of labelled data [3] and address the original challenge somewhat. Nonetheless, these systems are often limited by the quality of the data and the ability to generalise to new data [3]. This limitation causes the traditional expert system to be unable to catch up with its human counterpart as knowledge of the domain evolves.

The recent advancement of the GAI agent has helped to address some of these limitations by allowing the model to generate novel and unique ideas from existing data [3] such as text, image, sound and even video. This is a significant shift from the traditional expert system as it allows the model to be more flexible and adaptive to new data why bringing a new way of cooperating between humans and machines [3]. These GAI systems were quickly adopted in place of an expert to help users answer problems [20]. This is a unique property that emerges from the practical application of the technology despite its core nature, did not necessarily design the model to actually provide accurate answer for domain-specific questions.

This is especially the case for image generation, where the model is often used to generate a range of images that are used in the creative process [3]. However, this have mainly been focus evaluating text-based model and there is a gap in the research on evaluating image-based model. For the image-base model, researchers generally focused on the quality of the image generated and not the knowledge embedded within the image [39]. This in practice could lead to images that are visually appealing but lack the domain knowledge that is required for the image to be useful in the real world.

### 3.2 Criticism Regarding Quantitative Metrics for GAI

When evaluating deep learning models, statistical metrics are often used to evaluate their distribution and the error from the model’s desired target. However, for the DGNN model, this is often based upon the datatype that is the output such as text or image [30]. In practice, this means a fragmented set of metrics that are often not comparable.

In text base metrics, BLEU [29], ROUGE [24] are often used to evaluate the similarity between the generated text and the ground truth. However, these metrics often fail to capture the full extent of human knowledge and can fail when model input taxonomy relationships are complex such as in the case of a medical exam [26]. Thus, modern text-based evaluation metrics and datasets are still not robust enough to capture the full taxonomy relationship in human writing.

In image base metrics, FID [14], and Inception Score [6] are often used to evaluate the similarity between the generated image and the ground truth. However, critics of these metrics argue that they are often not reliably aligned to human perception [6]. These metrics are often based on the pixel-level similarity between the generated image and the ground truth. This means that the model can generate images that are visually similar to the ground truth but contain conflicting information such as a CT scan that contains both benign and malign features. In practice, this creates a seemingly coherent image at the pixel or local level but fails to capture the global structure of the image.

For an image model to be practical, it must present coherent visual and texture information, especially in high-precision fields like medical imaging, manufacturing, and electrical engineering, where model output directly impacts users [36]. Thus, our evaluation metric should capture both visual and texture details.

### 3.3 Criticism Regarding Qualitative Metrics for GAI

Since the literature shows quantitative metrics are often not good enough, modern research has started to look into qualitative metrics. Which often is the value of which the model is provided through human evaluation. This is often done through a range of methods such as human evaluation, expert evaluation, or user evaluation.

Researchers have tried to introduce image evaluation metrics such as Anomaly Score [15], and Aesthetic [25] that focus on capturing the naturalness and design function of the image. On text-based evaluations, researchers also introduced ConsiStory [33] such as which attempts to better capture semantic consistency in long text that is natural to humans such as stories. However, these metrics often do not factor in elements such as culture, education, and experience which contribute a large part of expertise within human [7]. Further research would be conducted to ensure these metrics represent the human aspect that they suggest and explore any edge case.

One major problem with image evaluation is the inconsistency in evaluation across domains. In domains like creative, images are largely evaluated on naturalness or creativeness [35]. But in highly technical fields such as engineering,

and manufacturing, images must follow a set of industry standards alongside a more concrete set of design principles [8]. This could lead to a model that is capable of performing well in one domain but not in another. Researchers have tried to introduce a range of methods to address this such as creating world model [12] which attempts to virtualise the domain environment, creating Holistic Benchmark that have an evaluation from different domain [1] or domain knowledge transfer method [36]. Hence, it might be unavoidable to say that the validation of the generated content might need to be domain-bound to ensure the system’s usability.

### 3.4 Limitation in the Generative Model Evaluation Methods

One such method to ensure that domain knowledge can be quickly evaluated is through the participation of human experts using Reinforcement Learning with Human Feedback (RLHF) [28]. Research has shown that human experts with deep knowledge of their field can often distinguish and identify problems in generated content [20] given enough time to interact with the model. However, one concern that arises is that the interaction between humans and models is often not easily distinguished from human to human using traditional application interfaces such as chatbox with limited GUI information [3]. The argument is that given enough training data, the model could learn to simulate the interaction signal that likely leads to a higher performance score, such as certain text or image patterns instead of the actual logical reasoning. Thus, the model might be more likely to trick the evaluator into giving it a high evaluation despite not performing the given tasks. Wolfert et al. [37] recent research shows that it is possible to mimic some behaviours of the end users to encourage trust in generative response regardless of the output performance.

One other aspect of interaction that could be overlooked is to understand how the model arrives at the decision it makes during the generation process. This can be extremely important when it involves a material generated in medical [18] or critical engineering environment such as industrial manufacturing [36]. Researchers [13, 26] have shown in their research that these DGNN models can often just memorise the test instead of understanding the questions present. Thus understanding how models make decision during their generative process might help to better inform developers and researchers to prevent current and future errors. Some researchers have attempted to address this with new techniques such as LIME [31] or counterfactual explanation based on gradual construction of the deep network [17] which show varied levels of success. Nonetheless, explainability is still among the most important areas of model-human interaction that should be considered during our evaluation process.

### 3.5 Knowledge Representation in Neural Network Architecture

Another challenge for our evaluation process is to embed human knowledge in machine learning and interpret the model decision-making process [3]. The most

common way to represent knowledge in machine learning tasks is directly embedding knowledge using high-dimensional vector or vector-space modelling [11]. Bordes et al. [5] show the versatility of vector space embedding compared to alternative symbolic frameworks.

Researchers [11, 5] generally agree that vector space embedding is the most appropriate technique to capture current knowledge in DGNN. However, vector space modelling is still not the end-all solution, while often able to capture semantic relationships in linguistic information, sensorimotor information (such as image or sound) is often not directly related to text information during training data collection [11]. In practice, this could mean that semantic information that is stored across different data types might not get represented in the embedded vector matrix.

One promising direction to address this problem is a Knowledge Graph (KG) to represent these data types and model their relationships through links between different edges that represent data types [36] before encoding information in vector form. Human knowledge is often interwoven with each other and is not discrete by nature. Utilising a graph as a data structure to represent domain knowledge offers the benefit of representing the complex relationship between different ontology in a domain and allows for a way to bridge different data types together [36]. Researchers [2] show that using KG alongside a KG-GAN help improve generative image quality by embedding additional knowledge that is implied in the text.

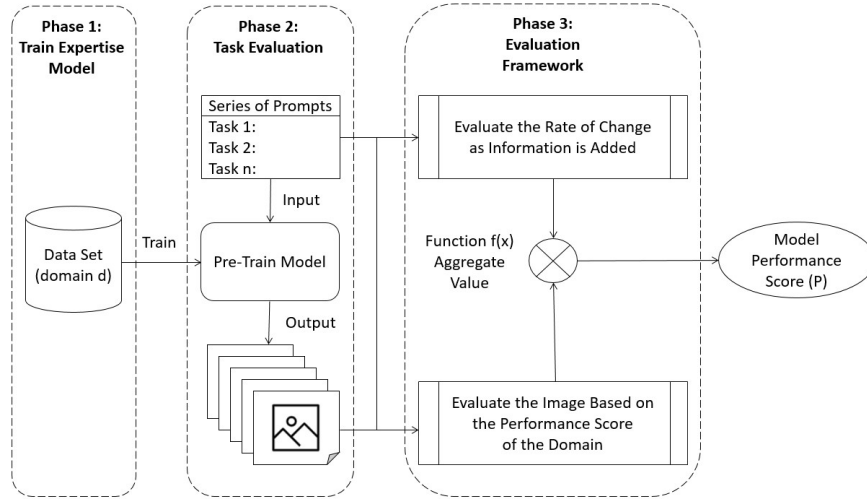
However, a criticism of the KG approach is that existing domain knowledge is not well represented and is often not complete [36]. This often means that existing data in these domains are not fully represented or potentially biased due to human interpretation [16]. In practice, this can introduce unwanted bias or reinforce existing misrepresentations of patterns in existing data.

## 4 Proposed Experiment Design to Evaluating Embedded Expert Knowledge of Generative Image Model

To address the existing challenge and attempt to verify our hypothesis, we propose the following research design which consists of different phases that are non-sequential and can be used in combination or individually to evaluate the model’s capabilities. Specifically, each model will be trained on a domain-specific dataset and subsequently tested on new semantic similar data. This approach ensures that the models do not merely memorise patterns from the pre-training data but genuinely learn from the additional training data provided.

The research design is outlined in Figure 1. We also wish to note that the research design is a high-level overview of the research process and each phase can be further broken down or modified to better suit the research needs. Thus, the research design is not a fixed process but a guideline that can be adopted for different industries and research goals.

The model will be evaluated on its ability to generate high-quality and domain-specific images based on the series of task prompts. We propose a se-



**Fig. 1.** Model capabilities evaluation research design

ries of prompts that are designed to test the model’s ability to generate images based on the updated information provided by each prompt. Essentially each prompt is build upon the previous prompt by either an increase in embedded information or the number of tokens. This idea is based upon the prompt chaining technique which shows promise in allowing for a control and understanding of the model output [38]. Wu et al. [38] research on prompt techniques shows that by chaining prompts together it is possible to further improve model reasoning without increasing the complexity of further tuning the model. Thus, we believe by adding this component to the research design we can overcome the need for large computational power while still reasonably challenging models’ reasoning capabilities. To further attempt to capture this change we propose a metric called "Rate of Change in Complexity" (RCC) which is calculated as:

$$RCC = g(TR, IR, \beta_0) \quad (1)$$

Where TR is the Complexity increase rate of the tasks, and IR is the Image Complexity Increase Rate. RCC is calculated as a function  $g()$  that takes three inputs TR, IR and  $\beta_0$ .  $\beta_0$  represents the baseline complexity ratio, the initial conditions or the inherent biases in task complexity relative to the image, independent of any specific changes in task or image complexity. In other words, this means that  $\beta_0$  accounts for any fixed, underlying complexity in the system that is not captured by the changes in task or image complexity rates.

For instance, if  $\beta_0$  is positive, it suggests that the task is inherently more complex relative to the image, even before any changes occur. Conversely, a negative  $\beta_0$  would indicate that the task is inherently less complex than the image, assuming no change in either complexity rate.  $\beta_0$  helps adjust the overall

complexity ratio RCC to better reflect real-world scenarios where tasks may have inherent complexities not solely attributable to changes in image complexity.

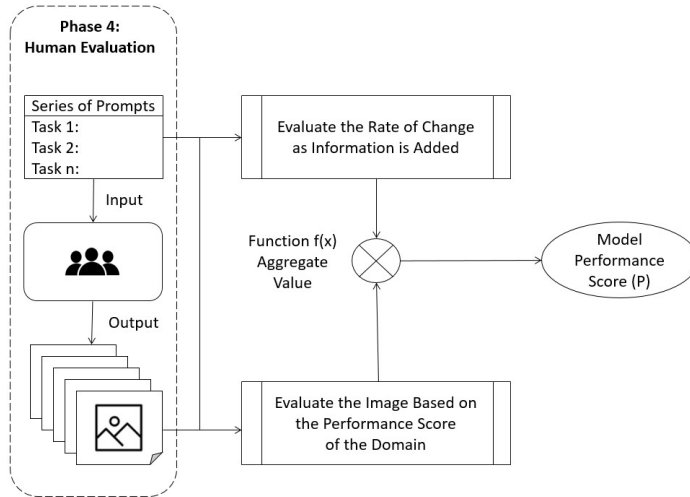
However, to ensure that the image serves the purpose of the task, we will also evaluate the image based on the domain-specific performance function. The performance function is determined by the domain expert and the best industry practice of the domain. The Performance Score is the mean of P across all T number of tasks with n being the number of different series of task T.

$$PerformanceScore = \frac{1}{n} \sum_{i=1}^n P_i \quad (2)$$

These two metrics are then combined to allow us to ensure the model meets the requirement of the task and can adapt to new information as the task complexity increases. This is approximate as  $\mathcal{P}$  with the function  $f()$  being a transformation function that regulates the value range.

$$\mathcal{P} \approx f(RCC, PerformanceScore) \quad (3)$$

As previously suggested by other researchers [4, 26] on the evaluation metrics performance of the model, likely, our metric will only be an approximation of the model performance compare to human expert and will need to be further refined in the future. Thus, we believe the value  $\mathcal{P}$  would be the value of an arbitrary approximation function and not a simple product of the RCC and Performance Score.



**Fig. 2.** Replacing our train model and repeating our test with the same set-up to compare our score with the group of human experts



As research regarding the performance of AI compared to humans has shown in most cases human experts can help to uncover unexpected flaws in our models [22]. Thus to evaluate if our model could be considered an expert in the domain, we will ask a group of human experts to evaluate the image generated by our model. In our research, The methodology involves a mixed-methods design, incorporating both between-subjects and within-subjects components [23] (see Figure 2). Specifically, a model and a group of human experts perform the same tasks, and their outputs are evaluated using standardised metrics. Subsequently, a separate group of human evaluators assesses these outputs without knowing their source, ensuring unbiased judgement. This also allowed for statistical evaluation and the alignment of our purpose metrics with the overall human evaluators.

## 5 Conclusion and Limitation

In our research proposition, we have found that the shift in the goals of evaluating AI has created a push toward more text-based evaluation of intelligence. This coupled with the constant growth and lack of agreed-upon evaluation metrics and tests has made it even more challenging to evaluate image generative model capabilities in capture and embedded knowledge. To address these challenges we propose a future research design in an attempt to explore ways to evaluate and train a model that can generate images that can satisfy high technical domain contain.

However, the paper also acknowledges some potential research design limitations and considerations. First limitation is in the access to existing computational power that enables the model to train sufficiently to address the task’s success. Future exploration could focus on compressing the tasks into a series of automated tests or libraries to meet lower computational requirements. Second, the scope of the task might not be complex enough for the domain space to properly evaluate the performance of the model compared to that of an expert. Future research could explore the model in terms of generating a sequence of actions such as assembling a guide or video to further explore how the model understands and retains these concepts continuously. Finally, the selected body of experts might not represent the latest and up-to-date of the domain. Future tests could be replicated across different bodies of experts with different levels of seniority across different locations to use collective intelligence.

By pointing out the existing challenge and proposing a framework for future research, the paper hopes to contribute toward the development of generative image technology into practical application in the high-constrained domain.

## Acknowledgement

The authors would like to thank Linh Nguyen, My Dinh and Dr Minh Dinh for their feedback and support during the writing of this paper.

## References

1. MMAU: A Holistic Benchmark of Agent Capabilities Across Diverse Domains
2. Ali, K.A.S.H., Krishna, S.C.: Generating text to realistic image using generative adversarial network. 2021 International Conference on Advances in Computing and Communications (ICACC) pp. 1–6 (2021), <https://api.semanticscholar.org/CorpusID:246870400>
3. Banh, L., Strobel, G.: Generative artificial intelligence **33**(1), 63. <https://doi.org/10.1007/s12525-023-00680-1>
4. Betzalel, E., Penso, C., Fetaya, E.: Evaluation Metrics for Generative Models: An Empirical Study **6**(3), 1531–1544. <https://doi.org/10.3390/make6030073>
5. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. Proceedings of the AAAI Conference on Artificial Intelligence (2011), <https://api.semanticscholar.org/CorpusID:715463>
6. Borji, A.: Pros and cons of GAN evaluation measures **179**, 41–65. <https://doi.org/10.1016/j.cviu.2018.10.009>
7. Ericsson, K.A.: Expertise and individual differences: The search for the structure and acquisition of experts’ superior performance **8**(1-2), e1382. <https://doi.org/10.1002/wcs.1382>
8. Fan, Z., Chen, T., Wang, P., Wang, Z.: Cadtransformer: Panoptic symbol spotting transformer for cad drawings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10986–10996 (2022)
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters **132**(2), 581–595
10. Goldman, A.I.: Expertise **37**(1), 3–10. <https://doi.org/10.1007/s11245-016-9410-3>
11. Günther, F., Rinaldi, L., Marelli, M.: Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. Perspectives on Psychological Science **14**, 1006 – 1033 (2019)
12. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering Diverse Domains through World Models, <http://arxiv.org/abs/2301.04104>
13. Heinz, M.V., Bhattacharya, S., Trudeau, B., Quist, R., Song, S.H., Lee, C.M., Jacobson, N.C.: Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health **9**
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. <https://api.semanticscholar.org/CorpusID:326772>
15. Hwang, J., Lee, J., Lee, J.S.: Anomaly Score: Evaluating Generative Models and Individual Generated Images based on Complexity and Vulnerability
16. Jain, N., Kalo, J.C., Balke, W.T., Krestel, R.: Do embeddings actually capture knowledge graph semantics? In: Extended Semantic Web Conference (2021)
17. Jung, H.G., Kang, S.H., Kim, H.D., Won, D.O., Lee, S.W.: Counterfactual explanation based on gradual construction for deep networks **132**, 108958–
18. Kather, J.N., Ghaffari Laleh, N., Foersch, S., Truhn, D.: Medical domain knowledge in domain-agnostic generative AI **5**(1), 90–90
19. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes, <http://arxiv.org/abs/1312.6114>
20. Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K., Lukowicz, P., Kuhn, J., Küchemann, S., Karolus, J.: Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education **386**

21. Kshetri, N.: Generative Artificial Intelligence in Marketing **25**(5), 71–75. <https://doi.org/10.1109/MITP.2023.3314325>
22. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., Chen, W.: ImagenHub: Standardizing the evaluation of conditional image generation models, <http://arxiv.org/abs/2310.01596>
23. Levy, B., Hilton, E.C., Tomko, M.E., Linsey, J.S.: Investigating problem similarity through study of between-subject and within-subject experiments (2017), <https://api.semanticscholar.org/CorpusID:65299281>
24. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, <https://aclanthology.org/W04-1013>
25. Liu, S., Xiang, Z., Yao, H., Cong, J.: A novel data-driven method for product aesthetics evaluating and optimising based on knowledge graph pp. 1–30
26. Lum, Z.C.: Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT **481**(8), 1623–1630. <https://doi.org/10.1097/CORR.0000000000002704>
27. Medsker, L.R.: Expert Systems and Neural Networks. In: Medsker, L.R. (ed.) Hybrid Intelligent Systems, pp. 39–56. Springer US
28. Moreira, I., Rivas, J., Cruz, F., Dazeley, R., Ayala, A., Fernandes, B.: Deep Reinforcement Learning with Interactive Feedback in a Human–Robot Environment **10**(16), 5574. <https://doi.org/10.3390/app10165574>
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. p. 311. Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>
30. Regenwetter, L., Srivastava, A., Gutfreund, D., Ahmed, F.: Beyond Statistical Similarity: Rethinking Metrics for Deep Generative Models in Engineering Design **165**. <https://doi.org/10.1016/j.cad.2023.103609>
31. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier, <http://arxiv.org/abs/1602.04938>
32. Soydaner, D., Wagemans, J.: Unveiling the factors of aesthetic preferences with explainable AI . <https://doi.org/10.1111/bjop.12707>
33. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-Free Consistent Text-to-Image Generation
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need
35. Wang, H.Y., Utama, S.: Investigating the Generative-AI Evaluation Methods and Correlation with Fashion Designers. In: 2023 7th International Conference on Information Technology (InCIT). pp. 508–513
36. Wang, L., Liu, J., Zhang, H., Zuo, F.: KMSA-Net: A Knowledge-Mining-Based Semantic-Aware Network for Cross-Domain Industrial Process Fault Diagnosis **20**(2), 2738–2750. <https://doi.org/10.1109/TII.2023.3296919>
37. Wolfert, P., Henter, G.E., Belpaeme, T.: Exploring the Effectiveness of Evaluation Practices for Computer-Generated Nonverbal Behaviour **14**(4), 1460–
38. Wu, T.S., Terry, M., Cai, C.J.: Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (2021)
39. Zhao, Z., Wang, S., Gu, J., Zhu, Y., Mei, L., Zhuang, Z., Cui, Z., Wang, Q., Shen, D.: Chatcad+: Towards a universal and reliable interactive cad using llms. IEEE transactions on medical imaging **PP** (2023)