Proceedings of the ASME 2024
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE 2024
August 25-28, 2024, August 25-28, 2024, Washington, DC, USA

IDETC2024-144076

CURRENT STATE AND BENCHMARKING OF GENERATIVE ARTIFICIAL INTELLIGENCE FOR ADDITIVE MANUFACTURING

Nowrin Akter Surovi[1,2], **Paul Witherell**[2], **Vinay Saji Mathew**[3], and **Soundar Kumara**[3]

[1]Singapore University of Technology and Design (SUTD)
[2]National Institute of Standards and Technology (NIST)
[3]The Pennsylvania State University

ABSTRACT

Additive Manufacturing (AM) is gaining popularity in the industry for its cost-effectiveness and time-saving benefits. However, AM encounters challenges that need to be addressed to enhance its efficiency. While Machine Learning (ML) can tackle various AM challenges, it is often limited to specific issues, necessitating multiple models. In contrast, Generative Artificial Intelligence (GenAI) has the potential to mitigate instance-specific bias due to its broader training. This paper presents a comprehensive methodology for evaluating the capabilities of various existing GenAI tools in addressing diverse AM-related tasks. We propose three categories of metrics, totaling 35 metrics, namely agnostic, domain task, and problem task metrics. Additionally, we introduce a scoring matrix, a practical tool that can be used to assess the responses of different GenAI tools. The study involves data collection from diverse published papers, which are used to create inquiries for GenAI tools. The results demonstrate that transformer-based models, such as multi-modal GPT-4 and Gemini (prev. BARD), can handle both AM image and text data. In contrast, uni-modals such as GPT-3 and Llama 2 are proficient in processing AM text data. Furthermore, image-based models such as DALL·E 3 and Stable Diffusion can accept AM text data and generate images. It is also observed that the performance of these models varies across different AM-related tasks. The variation in their performance may be due to their underlying architecture and the training dataset.

1 INTRODUCTION

Additive manufacturing (AM), also known as 3D printing, refers to a class of manufacturing process technologies associated with direct digital fabrication of complex geometrical objects from Computer-Aided Design (CAD) models using a layered manufacturing process. AM can be defined as “a process of joining materials to make objects from 3D model data, usually layer upon layer, as opposed to subtractive manufacturing methodologies” [1]. AM has several advantages over traditional manufacturing production techniques, including fabricating complex parts, achieving lightweight design, expediting and reducing production and delivery lead times [2]. Metal Additive Manufacturing (MAM) has gained prominence in various industries, particularly in sectors such as aerospace, defence, medicine, and energy, where unique challenges are met with unique solutions. As fabricated from digital manufacturing processes, AM-fabricated parts result in increasingly complex data streams from design to product transformation. These diverse data sets from AM processes contain valuable and actionable insights that can be used for deeper understanding and enhanced control of the AM process.

Machine Learning (ML) models play an important role in addressing various challenges within the design to product transformation in AM. The ML models, such as Neural Networks (NN), clustering, and Convolutional Neural Network (CNN) methods, are used for design [3,4,5], build precision, process parameter selection, optimization, part density prediction, [6,7] etc. ML models are also used in defect detection, process monitoring, and process control [8, 9, 10]. Moreover, ML models contribute to the study of dimensional variation classification [11, 12, 13]. The importance of ML in the AM domain is growing day by day. Figure 1 shows the increasing number of papers related to ML in AM. Most of the ML models applied in AM are domain and task-dependent. Each ML model is designed to handle specific issues, constituting a “bottom-up” approach. This approach implies the

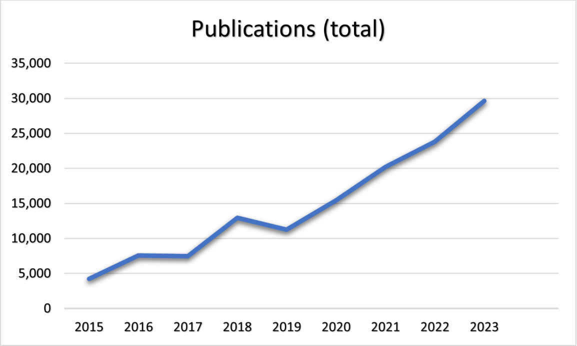

FIGURE 1: Number of ML papers in AM domain over time. Source: `https://app.dimensions.ai/`, Criteria: Machine learning in Additive Manufacturing

development of focused, domain-driven applications based on particular needs and opportunities. Consequently, different ML models are required to tackle different problems in AM. Moreover, most of the time, the ML models used in the AM domain are incapable of handling different modalities of data files. While some researchers have begun exploring multi-modal data handling in AM [14], these efforts remain limited, and these models are generally designed to tackle specific issues. Furthermore, most ML models available in the AM domain lack transparency regarding their details, such as training data sources, working environments, and real-world applications. Often, training data are inaccessible and non-reproducible. Therefore, utilizing published models with a different dataset may yield varying results due to variations in the dataset.

In such scenarios, Generative Artificial Intelligence (GenAI) emerges as a promising multi-task solution. GenAI is a more agnostic, top-down approach, where algorithms are initially trained on a broad range of data before narrowing a focus to a specific task. Because of its top-down approach, GenAI has the potential to improve upon single-task ML approaches by expanding solution spaces and reducing bias from focused training data. GenAI, through multi-modal tools, can process various data types (e.g., images, videos, and acoustic signals) and solve various problems simultaneously. Thus, it eliminates the need for multiple specialized models, as a single GenAI tool can potentially handle different modalities of data and address various issues related to AM at the same time.

To evaluate the effectiveness of existing GenAI tools in addressing MAM tasks, three types of benchmarking task metrics: agnostic, domain task, and problem task, totaling 35 metrics, are proposed. These metrics are selected based on various GenAI opportunities or dimensions within the four exploration spaces (Figure 2). In this paper, we evaluate six popular GenAI tools, namely GPT-4, GPT-3.5, Gemini (prev. BARD), Llama 2, DALL·E 3, and Stable Diffusion. The primary contribution is proposing an initial set of metrics on which benchmarking can be performed. We also propose a scoring matrix to quantify the performance of each tool. To score these metrics, we develop a variety of text-based and image-based prompts based on published AM-related literature for the GenAI tools. We then assess the responses generated by these tools and benchmark their performance based on the obtained scores.

2 BACKGROUND AND LITERATURE REVIEW

In this section, we will discuss GenAI and its subdivision based on different criteria and 6 different popular GenAI models.

2.1 Generative Artificial Intelligence (GenAI) and Its Classification

Generative Artificial Intelligence (GenAI) refers to algorithms capable of generating novel, creative and realistic content, including images, audio, video, and 3D models, replicating real data distributions [16]. In exploring and bench-marking GenAI, in this paper, we categorize GenAI based on the architecture shown in Figure 3 and modality shown in Figure 4.

The classification of GenAI models based on their architecture provides insights into their fundamental components and training methods. For instance, Variational Autoencoders (VAEs) adopt an encoder-decoder architecture and employ variational inference during training. Generative adversarial networks (GANs) leverage adversarial training, featuring a generator and discriminator for creating realistic and diverse data. Diffusion models involve a noising and denoising process, iteratively refining noisy inputs for high-quality samples. Transformers, with encoder-decoder architecture and self-attention mechanisms, capture global dependencies through supervised training. Language models, often based on recurrent neural networks (RNNs), generate natural language sequences by predicting the next token through supervised learning. Normalizing flow models use coupling layers for data transformation while preserving density and learning complex distributions. Hybrid models combine various architectures and training methods by integrating elements from multiple models [17].

The classification of GenAI models based on modality provides insights into their ability to process specific data types like text, images, audio, or video. Uni-modal models generate results in the same format as the input prompts. For example, GPT-3.5 utilizes text-to-text generation, and GAN, VAE, and Normalizing Flow utilize image-to-image generation. Multi-modal models can process prompts from various modalities and generate results in multiple modalities. These models handle both input and output of different modalities (e.g., image-to-text) or multi-modal inputs (e.g., processing both text and images) and outputs (e.g., generating both text and images). Examples of multi-modal models include the use of DALL·E and VisualBERT for text-to-image generation, AdaSpeech for text-to-audio, KG-BERT for

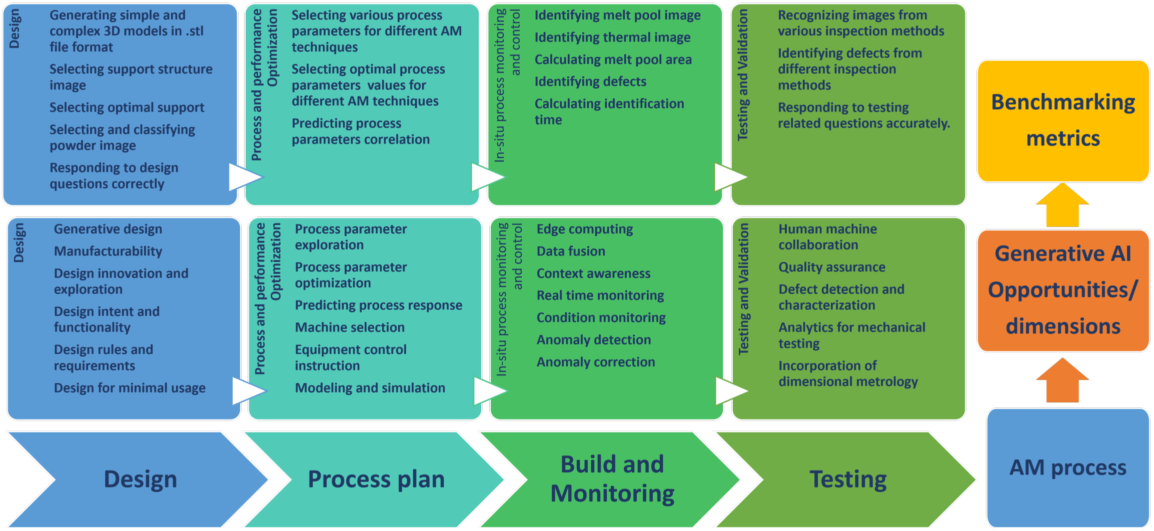
Design
Generating simple and complex 3D models in .stl file format
Selecting support structure image
Selecting optimal support
Selecting and classifying powder image
Responding to design questions correctly
Process and performance Optimization
Selecting various process parameters for different AM techniques
Selecting optimal process parameters values for different AM techniques
Predicting process parameters correlation
In-situ process monitoring and control
Identifying melt pool image
Identifying thermal image
Calculating melt pool area
Identifying defects
Calculating identification time
Testing and Validation
Recognizing images from various inspection methods
Identifying defects from different inspection methods
Responding to testing related questions accurately.
Benchmarking metrics
Design
Generative design
Manufacturability
Design innovation and exploration
Design intent and functionality
Design rules and requirements
Design for minimal usage
Process and performance Optimization
Process parameter exploration
Process parameter optimization
Predicting process response
Machine selection
Equipment control instruction
Modeling and simulation
In-situ process monitoring and control
Edge computing
Data fusion
Context awareness
Real time monitoring
Condition monitoring
Anomaly detection
Anomaly correction
Testing and Validation
Human machine collaboration
Quality assurance
Defect detection and characterization
Analytics for mechanical testing
Incorporation of dimensional metrology
Generative AI Opportunities/ dimensions
Design
Process plan
Build and Monitoring
Testing
AM process

FIGURE 2: Digital flow of AM and Generative AI dimensions in each AM phase, with selected metrics under these dimensions. Adapted from [15]

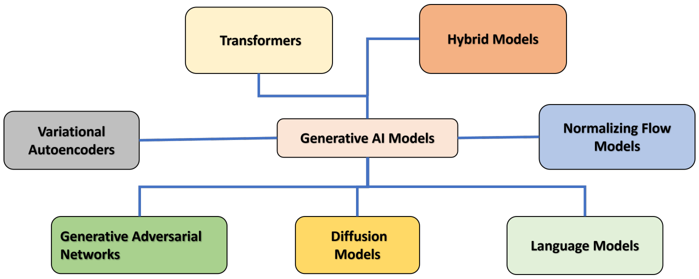

FIGURE 3: Classification of GenAI based on Architecture

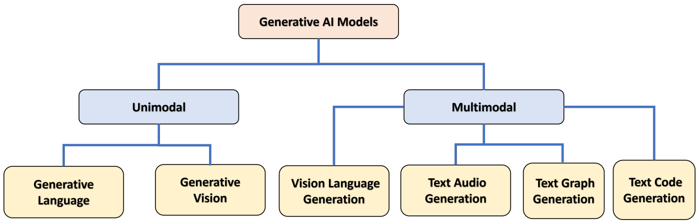

FIGURE 4: Classification of GenAI based on Modality

text-graph processing, and CodeBERT and CodeX for text-to-code generation [18].

2.2 Current State of GenAI in AM

Different GenAI models are currently utilized to address diverse challenges across different phases of AM.

In the design domain, NASA GSFC developed a lightweight generative design process to demonstrate potential savings in development time and mass [19]. Elbadawi et al. [20] utilized conditional generative adversarial networks (cGANs) to facilitate Fused Diffusion Model (FDM) printing. For topology optimization, Hertlein et al. [21] developed a cGAN-based framework to predict optimal designs for AM without overhangs. In monitoring and control, Petrik et al. [22] introduced MeltPoolGAN for classifying melt pool images and optimizing process parameters. Mu et al. [23] developed an adaptive online simulation model using a diffusion-based Generative AI model and laser-scanned point clouds to predict distortion fields in new deposition cases. The model served as a foundation for model-based control systems, topology optimizations, and advancements in metallic additive manufacturing design and technology (AM-DTs). Liu et al. [24] developed an image-enhancement generative adversarial network (IEGAN) to improve the quality of thermal images for image segmentation.

Beyond the facilitation of the design-to-product transformation, researchers also leverage GenAI tools for studying AM software and fundamentals. Badini et al. [25] assessed the capability of GPT for optimizing G-code and printing parameters in Fused Filament Fabrication (FFF) AM. Jignasu et al. [26] used six GenAI tools to comprehend and debug G-code files for 3D printing. Fang et al. [27] employed ChatGPT and BERT to enhance the accuracy of a graph for recycled metal powder.

2.3 GenAI Models Examples

In this section, we provide short descriptions of some of the most prominent GenAI models.

2.3.1 GPT-4 & GPT-4V GPT-4 (Generative Pre-trained Transformer 4) [28] is a state-of-the-art transformer-based lan-

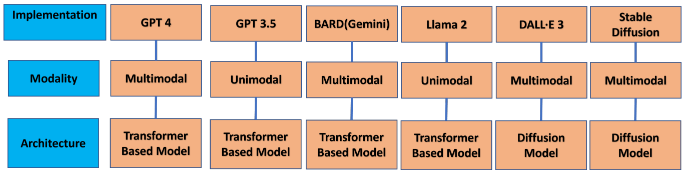

FIGURE 5: Combined classification of GenAI tools

guage model trained on a large amount of text and image data. It employs a transformer architecture to generate human-like text based on its input. An extension of GPT-4, or GPT-4V, incorporates vision capabilities [29], enabling it to process and generate responses based on textual and visual modalities. Therefore, the GPT-4 model can handle AM text and image data.

2.3.2 GPT-3.5 GPT-3.5 (Generative Pre-trained Transformer 3.5) is a language model developed by OpenAI, which serves as precursor [30] to GPT-4. GPT-3.5 has been trained on a larger dataset and features several improvements over its predecessors. Unlike GPT-4, GPT-3.5 has no integrated vision capabilities and focuses solely on processing and generating text. Therefore, GPT-3.5 can handle only AM text datasets.

2.3.3 Gemini (Previously Bard) Google Gemini, previously known as Bard, is a conversational AI model built on the foundation of LaMDA (Language Model for Dialogue Applications) [31]. It's unique in its ability to process both images and text as input and generate text as output. This versatility enables it to effectively handle both AM image and text data.

2.3.4 Llama 2 (Large Language Model Meta AI) Llama 2 is an updated version of Meta AI's original Llama model [32]. It is open access for research and commercial use. It can only handle AM text data.

2.3.5 DALL·E DALL·E, an AI model by OpenAI [33], is a generative model that produces images from textual descriptions. As a result, it's capable of generating AM image data based on textual input.

2.3.6 Stable Diffusion (SD) Stable Diffusion, a latent text-to-image diffusion model, is a collaborative development by Stability AI, Runway, and CompVis [34]. SD can generate AM image files based on textual input.

The combined classification of the above-mentioned GenAI models based on architecture and modality is shown in Figure 5.

3 METRICS FOR BENCHMARKING GENAI TOOLS

In this section, we introduce three distinct types of metrics based on the GenAI opportunities or dimensions across the four phases of the AM process (Figure 2) for benchmarking existing GenAI tools. The selection of metrics is primarily guided by the complexity of AM tasks across these phases: Design, Process Plan, Build and Monitoring, and Testing. These four phases are a simplified version of the eight phases proposed by Kim et al. [35].

1. Agnostic Metrics: These are characterized by their independence from any particular AM phase or task. They offer a broad perspective on overall performance without being tied to specific processes or stages.
2. Domain Task Metrics: These refer to the generic tasks or activities directly related to the specific domain or phase within AM. They are independent of a specific problem but depend on certain AM phases.
3. Problem Task Metrics: These refer to the challenges or issues that arise within the specific AM domain or phase requiring problem-solving skills. These tasks are generally more complex and specific than domain tasks. They depend on both AM phases and particular problems.

A scoring matrix is also introduced for all metrics to evaluate responses from various GenAI tools. Therefore, the metrics and scoring matrix create a robust and reliable approach for assessing the effectiveness of GenAI tools in AM.

The following section explains all three distinct types of metrics and their corresponding scoring metrics.

3.1 Agnostic Metrics

Given the general applicability of many GenAI tools, the agnostic performance metrics were an important starting point for providing a baseline capability evaluation. While AM has several distinct phases, many of the problem types remain the same, particularly in the context of the exploration of solution space and the data processing requirements. Except for two closely related ones, each of the tools here is based on different GenAI models with approaches to tasks and different adaptations of architectures. To establish a baseline of tool capability, five separate performance metrics were selected: The number of supported input data types, number of supported output data types, data compatibility ratio, response time for text, and response time for image generation.

These five metrics were chosen to give general insight into the basic utility of the different GenAI tools, including assessing how well they support different types of data and what relative response times might be. These metrics were selected to provide insight into the general capabilities of the GenAI tools before a deeper dive is performed in the domain-specific areas.

3.2 Domain Task Metrics

The domain-specific metrics across the four distinct AM phases were important for evaluating GenAI tools. In this section, fifteen domain task metrics have been suggested.

3.2.1 Design For the Design phase, four metrics (Table 2) were introduced to evaluate the performance of GenAI tools. These metrics were chosen to offer insights into how well

TABLE 1: Agnostic metrics with scoring matrix

Metrics	Scores				
	5	4	3	2	1
Number of supported input data types	$5 \leq$	4	3	2	1
Number of supported output data types	$5\leq$	4	3	2	1
Data compatibility ratio	$5\leq$	4	3	2	1
Response time for text	≤ 1 s	≤ 5 s	≤ 10 s	≤ 30 s	30 s$\leq$
Response time for image generation	≤ 1 s	≤ 5 s	≤ 10 s	≤ 30 s	30 s$\leq$

GenAI tools perform in design domain-specific tasks. These include their ability to generate 3D models, recognize powder images used in AM, answer design-related questions, and identify support structure images, etc.

As would be expected, much of the design phase evaluation focuses on assessing a tool's ability to interpret and manipulate geometry. The metrics chosen for the design phase were meant to investigate some specific challenges of the design phase, such as the ability to process geometry for topological optimization, the ability to differentiate between similar but different shapes as might be encountered in feedstock evaluation, and the ability to distinguish between similar but different geometries, as might be encountered in support structure development.

TABLE 2: Domain task metrics: Design and their corresponding scoring matrix

Metrics	Scores				
	5	4	3	2	1
Generate 3D Model	3D model in chosen format	Model instructions	Incomplete design	3D image	Unable to generate
Identify powder image	Identify as powder image	Identify similar image type	Contextualization, no identification	Incorrect Contextualization	No Context, Unable to identify
Respond to design questions	Respond to all questions	Respond one less than all questions	Respond two less than all questions	Respond to less than half questions	Unable to respond
Identify support structure from image	Able to identify	Contextualization, no identification	Partially contextualization, no identification	Incorrect Contextualization	No context, Unable to identify

3.2.2 Process Plan

For the Process Plan phase, four metrics (Table 3) were chosen to evaluate GenAI tools. The focus of choosing these metrics was on assessing the performance of GenAI tools for selecting the number of process parameters (PP) for Powder Bed Fusion (PBF), Selective Laser Melting (SLM), and Wire Arc Additive Manufacturing (WAAM) processes. Additionally, an assessment was made to determine whether GenAI tools can predict the interrelation among various process parameters in the PBF process.

The process planning and processing stages are the two most technology-dependent phases in the AM life cycle. Subsequently, during these phases, the data types and problems presented may differ significantly, and some insight into those potential differences is important. Perhaps more than any other stage, the process planning stage benefits from simulation and exploration of parameter settings; thus, the ability to explore large parameter configurations and simulated time series data may be important.

TABLE 3: Domain task metrics: Process Plan and their corresponding scoring matrix

Metrics	Scores				
	5	4	3	2	1
Select PBF parameter	Select maximum number	Select one under max	Select two under max	Select three or more under max	Unable to select
Select SLM parameter	Select maximum number	Select one under max	Select two under max	Select three or more under max	Unable to select
Select WAAM parameter	Select maximum number	Select one under max	Select two under max	Select three or more under max	Unable to select
Identify process parameter relationships	Select maximum number	Select one under max	Select two under max	Select three or more under max	Unable to select

3.2.3 Build and Monitoring

For the build and monitoring phase, three metrics (Table 4) were suggested for assessing the GenAI tool's capability, including the identification of melt pools and thermal images and the response time of the identification for real-time monitoring tasks.

More than the other three phases, the time element was emphasized in building and monitoring metrics, specifically to gauge the potential for these tools to be used in real or near real-time responses. As the building and monitoring phases also rely on large amounts of time series sensor data, this evaluation also emphasized the ability to handle time series data and different data types.

TABLE 4: Domain task metrics: Build and Monitoring and their corresponding scoring matrix

Metrics	Scores				
	5	4	3	2	1
Identify melt pool image	Identify melt pool image	Recognize a similar form of image	Contextualization, no identification	Incorrect Context, no identification	Unable to select
Identify thermal image	Identify thermal image	Recognize a similar form of image	Contextualization, no identification	Incorrect Context, no identification	Unable to select
Average Identification time	≤ 1 s	≤ 5 s	≤ 10 s	≤ 30 s	30 s$\leq$

3.2.4 Testing

For the Testing phase, four metrics (Table 5) were proposed to give insight into the utility of the GenAI tools in recognizing scanning electron microscopy (SEM), high-resolution camera, and X-CT images, along with their ability to respond to post-processing and testing-related questions.

Tasks in the testing phase will focus on identifying abnormalities or defects in parts through techniques such as optical measurements or XCT. Such tasks will likely benefit from the ability to identify and operate on thresholds. Another challenge at the testing level is the ability to interpret and differentiate between two-dimensional and three-dimensional geometries. This phase may rely on statistical data more than any other phase due to the need to assess material properties.

3.3 Problem Task Metrics

The problem-specific metrics across the four distinct AM phases were important for evaluating GenAI tools. In this section, fifteen problem task metrics have been proposed.

3.3.1 Design

For the Design phase, four metrics (Table 6) were proposed to determine whether GenAI tools can generate complex 3D models with specific measurements, classify different powders and select optimal support structures based on their images. Additionally, an investigation was conducted into whether various design-related questions can be responded to accurately by the GenAI tools.

Generating 3D models with specific measurements is important for producing manufactured parts that meet specific requirements accurately and precisely. Again, the classification of powder aids in comprehending the properties, quality, and characteristics of the final printed parts. Furthermore, carefully selecting optimal support structures is essential to minimize material waste during printing.

TABLE 5: Domain task metrics: Testing and their corresponding scoring matrix

Metrics	Scores				
	5	4	3	2	1
Identify SEM image	Identify as SEM image	Recognize a similar form of image	Contextualization, no identification	Incorrect Contextualization	No context, Unable to identify
Identify high-resolution camera image	Identify as high-resolution camera image	Recognize a similar form of image	Contextualization, no identification	Incorrect Contextualization	No context, Unable to identify
Identify X-CT image	Identify as X-CT image	Recognize a similar form of image	Contextualization, no identification	Incorrect Contextualization	No context, Unable to identify
Respond to testing questions	Respond to all questions	Respond one less than all questions	Respond two less than all questions	Respond to less than half questions	Unable to respond

3.3.2 Process Plan

For the Process Plan phase, four metrics (Table 7) were proposed to evaluate the performance of GenAI tools for selecting suitable process parameter ranges for PBF, SLM, and WAAM processes. Assessment was also made regarding accurately identifying relationships among different process parameters for PBF.

Since each AM technique has unique characteristics, material requirements, and process dynamics, it is important to select an appropriate process range for each AM method. This selection enables control over factors like melting and solidification rates, defect minimization, and maintaining build speed.

3.3.3 Build and Monitoring

For the Build and Monitoring phase, three metrics (Table 8) were proposed to determine how well the GenAI tools calculate melt pool area and detect anomalies from melt pool and thermal images.

Defect detection in real-time saves material resources and allows for immediate corrective actions during the printing process. Calculating the melt pool area is also important as it reflects the current state of the AM process and is directly related to the quality of the final parts.

TABLE 6: Problem task metrics: Design and their corresponding scoring matrix

Metrics	Score				
	5	4	3	2	1
Generate dimensioned 3D model	3D model for chosen format	Model instruction	Incomplete design	3D image	Unable to generate
Classify AM powder from images	Able to classify	Provide hint, no exact classification	Contextualization, no classification	Incorrect Contextualization, no classification	Unable to classify
Number of correct answers	Correctly answers all questions	One incorrect answer	Two incorrect answer	Three incorrect answer	Unable to correct answer
Select optimal support from image	Able to select that matches with reference	Provides hint, no exact selection	Contextualization, no selection	Incorrect Contextualization, no selection	Unable to select

3.3.4 Testing

For the Testing phase, four metrics (Table 9) were proposed to give insight into the utility of the GenAI tools to determine defects from SEM, high-resolution camera images, and porosity labels from X-CT images. Additionally, an assessment was made regarding how well these GenAI tools address testing-related questions.

The identification of defects from different sources during testing phases is essential for evaluating the integrity, performance, and reliability of the final product. Additionally, diverse testing-related information, methods for enhancing parts, and equipment are crucial for assessing material properties.

4 RESULTS AND DISCUSSION

We collected images and information data from published papers to create prompts [36, 37, 38, 9, 39, 40, 41, 42, 43]. We used these prompts to generate responses five times for each GenAI tool using their API because these tools tend to give different styles of responses each time. Hence, we reviewed these responses to limit variability and chose the one closest to the reference. Based on the response from these tools, we scored each GenAI tool within each metric. This scoring helps us to benchmark the tools by comparing the scores. All questions and responses for scoring agnostic, domain task and problem task metrics are available on GitHub: `https://github.com/nowrin0102/IDETC-2024`. Note that we maintain the same prompt for all the models for fair comparisons.

TABLE 7: Problem task metrics: Process Plan and their corresponding scoring matrix

Metrics	Score				
	5	4	3	2	1
Select laser power & scan speed for PBF	Select both exact as reference	Select both close to reference	Select one close to reference	Contextualization, no selection	Unable to select
Select laser power & scan speed for SLM	Select both exact as reference	Select both close to reference	Select one close to reference	Contextualization, no selection	Unable to select
Select torch speed & wire feed rate for WAAM	Select both exact as reference	Select both close to reference	Select one close to reference	Contextualization, no selection	Unable to select
Predict process parameter correlations	Predict exact as reference	Wrongly predicts one	Wrongly predicts two	Wrongly predicts at least three	Unable to select

TABLE 8: Problem task metrics: Build and monitoring and their corresponding scoring matrix

Metrics	Score				
	5	4	3	2	1
Calculate melt pool area from image	Able to calculate	Able to calculate, close to actual value	Existence of melt pool	Contextualization, no detection	Unable to identify
Detect anomaly from image	Able to detect	Able to detect partially	Existence of melt pool	Contextualization, no identification	Unable to identify
Identify defect from thermal image	Able to identify	Recognise a similar form	Contextualization, no identification	Incorrect Contextualization, no identification	Unable to identify

4.1 Agnostic Metrics Results

In Figure 6, scores for various GenAI tools on agnostic metrics are summarized. Firstly, GPT-4 and Gemini support a more diverse range of inputs (images, text, formulas, code, and mathematical expressions) compared to GPT 3.5 and Llama 2, which

TABLE 9: Problem task metrics: Testing and their corresponding scoring matrix

Metrics	Scores				
	5	4	3	2	1
Identify defect from SEM image	Able to identify	Recognize a similar form	Contextualization only	Incorrect Contextualization	No Context, Unable to identify
Identify defective part from image	Able to identify	Recognize a similar form	Contextualization only	Incorrect Contextualization	No Context, Unable to identify
Identify porosity from X-CT image	Able to identify	Partial identification	Contextualization only	Incorrect Contextualization	No Context, Unable to identify
Understand testing context	Correctly answer all questions	One incorrect answer	Two incorrect answers	Three or more incorrect answers	Unable to correct answer

cannot process image data. Furthermore, all four tools can generate text, formulas, code, and mathematical data used in AM as output. Secondly, GPT-4 and Gemini can simultaneously handle images and text data, providing a 2:1 data compatibility ratio. Thirdly, GPT 3.5 exhibits faster responses than GPT-4 and Gemini. Additionally, DALL·E 3 consistently takes longer to generate images than Stable Diffusion. It's important to note that we calculate response time in real-time using their API. Since the response time may vary based on query size, bandwidth, server load, etc., we provide scores instead of actual time duration.

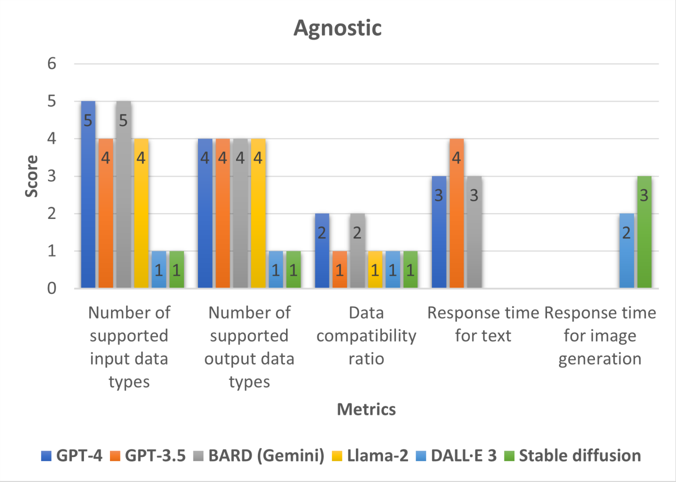

FIGURE 6: Agnostics Metrics Evaluation

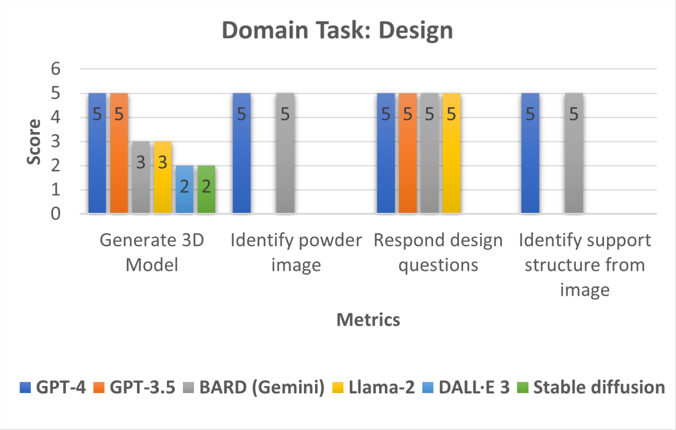

FIGURE 7: Domain task Evaluation: Design

4.2 Domain Task Metrics Results

In the Design results (Figure 7), it is observed that GPT-4 and GPT-3.5 can generate 3D models (.stl/OpenSCAD format files), earning a score of 5. Gemini and Llama 2, though attempting 3D file generation, often produce incomplete designs. The proficiency of GPT-4 and GPT-3.5 in 3D model generation may be attributed to diverse training datasets related to 3D printing and design. Despite sharing a transformer-based architecture, differences in the models' structures could impact 3D model generation. DALL·E 3 and stable diffusion, being diffusion-based models, can generate images of the specified objects but are limited to image generation. GPT-4 and Gemini, as multi-modal models, can handle both image and text inputs and can identify powder and support structures from images accurately. Notably, all transformer-based models display proficiency in answering design-related questions due to their text input processing capabilities.

In the Process Plan results (Figure 8), Gemini and Llama2 can select the maximum number of process parameters for PBF and WAAM, so they score 5. Additionally, Gemini achieves the maximum score for selecting process parameters for SLM. This is likely due to their extensive training on diverse process-related datasets and ability to generalize to queries related to process parameters of the AM process. Llama 2 can establish more relationships between different process parameters for PBF than others, indicating its understanding of these interconnections.

In the Build and Monitoring results (Figure 9), it is observed that multi-modal models, GPT-4 and Gemini, can process the melt pool and thermal image data. GPT-4 is able to identify both melt pool and thermal images and achieve the maximum score, while Gemini contextualizes the melt pool data but struggles to identify thermal images. This could be attributed to either differ-

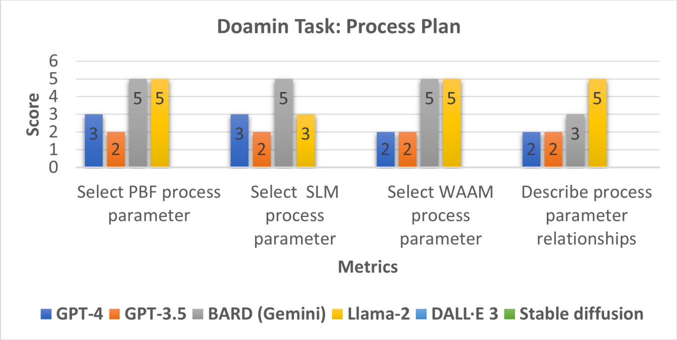

FIGURE 8: Domain Task Evaluation: Process Plan

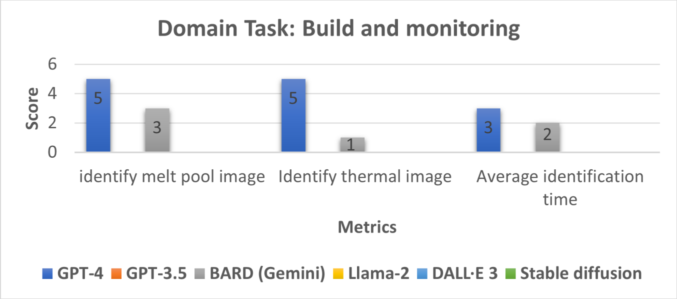

FIGURE 9: Domain Task Evaluation: Build and monitoring

ences in their training datasets, underlying architecture or model size. It is also observed that the identification response time is consistently faster for GPT 4, although, again, the response time depends on the query size, bandwidth, server load, etc.

In Testing results (Figure 10), the multi-modal models GPT-4 and Gemini are able to handle SEM, high-resolution camera and X-CT images. GPT-4 can identify SEM and high-resolution camera images, while Gemini can identify SEM images and contextualize them for high-resolution camera images. Moreover, both GPT-4 and Gemini can contextualise the X-CT image data. It is also observed that all the transformer-based models can answer testing-related queries as they can handle text input.

4.3 Problem Task Metrics Results

In the Design results (Figure 11), it is observed that all tools frequently generate incomplete or misleading designs, offering suggestions on how to proceed. This may be due to their insufficient competence in generating complex 3D models, attributed to the limitations of their training datasets and inherent architecture. Although Gemini and GPT-4 can handle image data, they show differences in their performance when classifying powder. Gemini can provide hints for classifying powder from images but doesn't provide exact classification, while GPT-4 lacks this capability. GPT-4 and Gemini both offer hints about potential support structure but cannot provide exact answers close to the reference paper [44]. GPT-4 and Gemini can respond to design-related questions closer to the reference than GPT-3.5 and Llama 2, yet none of the models can provide all correct answers. This may be due to GPT-4 and Gemini having more extensive training on design-related datasets than 3.5 and Llama 2.

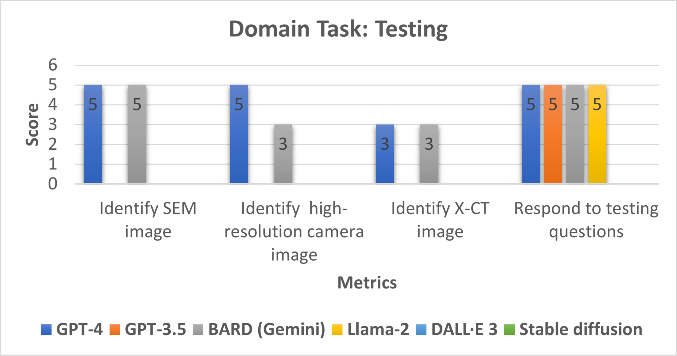

FIGURE 10: Domain Task Evaluation: Testing

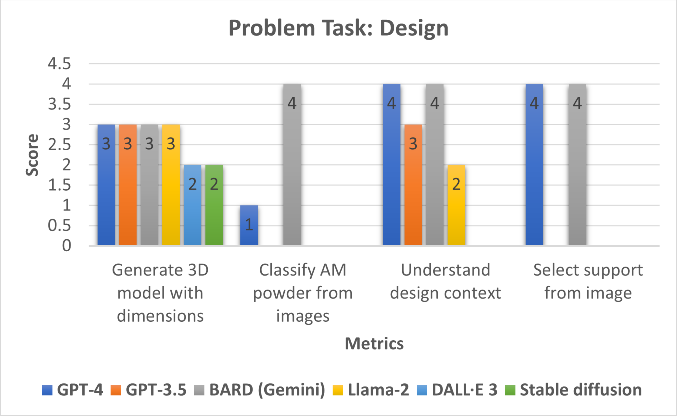

FIGURE 11: Problem Task Evaluation: Design

In the Process Plan results (Figure 12), all transformer-based models can predict the range of process parameters and the relationships among different parameters. GPT-4 and Gemini consistently predict process parameter ranges close to the reference compared to GPT 3.5 and Llama 2.

In Build and Monitoring results (Figure 13), both GPT-4 and Gemini calculate melt pool area, though not precisely matching the reference values. They can identify defective melt pool images based on a detailed prompt description. Gemini can explain possible defects in thermal images, while GPT-4 shows incapability to identify anything.

In Testing results (Figure 14), none of the models can precisely identify defects from SEM images. GPT-4 guesses possible defects in SEM and XCT images and scores 4, while Gemini

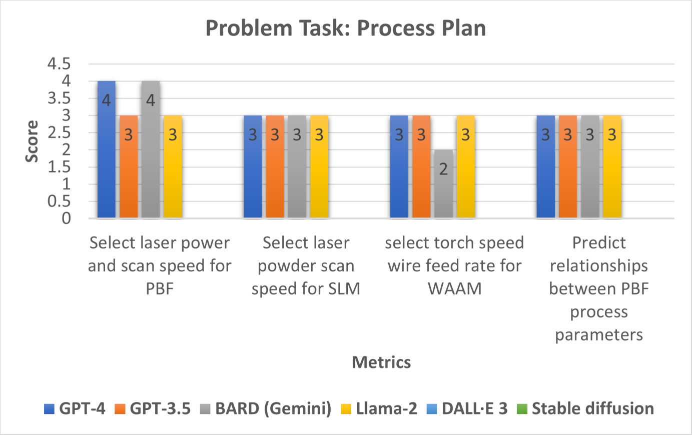

FIGURE 12: Problem Task Evaluation: Process Plan

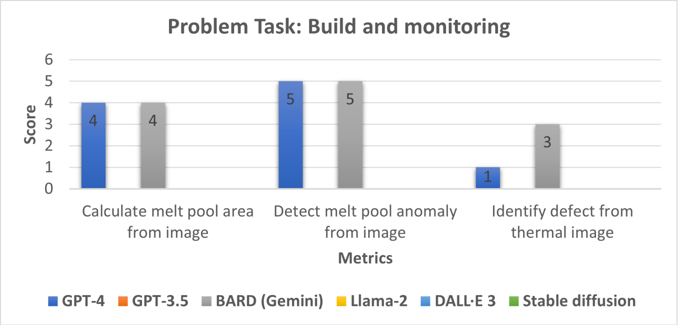

FIGURE 13: Problem Task Evaluation: Build and Monitoring

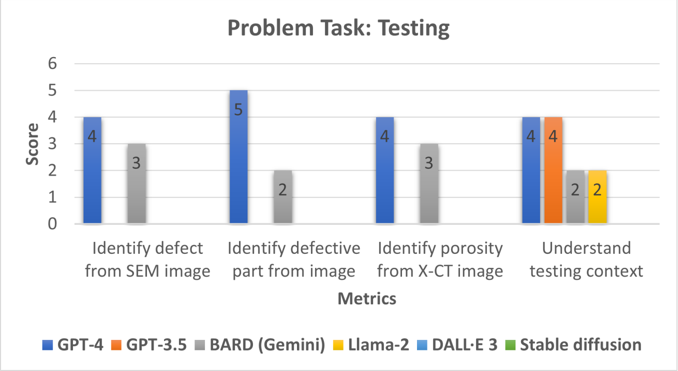

FIGURE 14: Problem Task Evaluation: Testing

provides explanations. Additionally, unlike Gemini, GPT-4 can identify defects from high-resolution camera images. GPT-4 and GPT-3.5 deliver responses to testing-related questions closer to the reference than Gemini and Llama 2, likely because of the commonalities GPT 4 and GPT 3.5 share with respect to their training corpus.

In summary, GenAI presents numerous opportunities to address diverse AM tasks. The metrics are chosen based on the four phases of AM and the corresponding GenAI opportunities. After evaluating GenAI tools using these metrics, we conclude that all existing GenAI models show competence in handling various domain-related tasks. However, they have limitations in solving specific problem tasks. Performance variations are likely due to differences in modality, architecture, training datasets, and the number of model parameters.

5 CONCLUSION

In this study, we have introduced three categories of metrics based on four AM exploration spaces: agnostic, domain task, and problem task metrics, totaling 35 metrics. These metrics are used to evaluate the capabilities of six popular existing GenAI tools. The selected GenAI tools include GPT-4, GPT-3.5, Gemini (formerly BARD), Llama 2, DALL·E 3, and Stable Diffusion. We have also proposed a scoring matrix to assess the responses of these GenAI tools. By utilizing data from published papers, we have created inquiries, evaluated responses, and assigned scores based on the proposed scoring matrix. After comparing the scores across various metrics, we have found that different tools have different processing capabilities. We have also noticed that most of these existing models perform well for domain task metrics; their performance in tackling specific problem tasks is less consistent. The performance variation may be attributed to the underlying architecture of the models and their training dataset. We've outlined our future tasks in three parts. Firstly, the metrics selected under GenAI dimensions in this paper do not cover all AM-related tasks. In the future, we will broaden our metric selection to encompass all relevant AM tasks. Secondly, the current number of questions used to evaluate GenAI tools is limited. We plan to increase the number of queries to make our benchmarking more robust. Thirdly, we are working on developing a customized model specifically designed to solve complex problem-solving tasks.

6 Acknowledgements

This research was funded by the National Institute of Standards and Technology (NIST), U.S. Department of Commerce, under the Additive Manufacturing Program. Certain commercial and third-party products are identified in this paper. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products identified are necessarily the best available for the purpose.

REFERENCES

[1] Standard, A., et al., 2012. "Standard terminology for additive manufacturing technologies". *ASTM International F2792-12a*, pp. 1–9.

[2] Frazier, W. E., 2014. "Metal additive manufacturing: a review". *Journal of Materials Engineering and performance*, ***23***, pp. 1917–1928.

[3] Yao, X., Moon, S. K., and Bi, G., 2017. "A hybrid machine learning approach for additive manufacturing design feature recommendation". *Rapid Prototyping Journal*, ***23***(6), pp. 983–997.

[4] Chan, S. L., Lu, Y., and Wang, Y., 2018. "Data-driven cost estimation for

additive manufacturing in cybermanufacturing". *Journal of manufacturing systems,* ***46***, pp. 115–126.

[5] Gaynor, A. T., 2015. "Topology optimization algorithms for additive manufacturing". PhD thesis, Johns Hopkins University.

[6] Fathi, A., and Mozaffari, A., 2014. "Vector optimization of laser solid freeform fabrication system using a hierarchical mutable smart bee-fuzzy inference system and hybrid nsga-ii/self-organizing map". *Journal of Intelligent Manufacturing,* ***25***, pp. 775–795.

[7] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., and Lu, Y., 2018. "A super-metamodeling framework to optimize system predictability". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 51722, American Society of Mechanical Engineers, p. V01AT02A009.

[8] Surovi, N. A., and Soh, G. S., 2023. "Acoustic feature based geometric defect identification in wire arc additive manufacturing". *Virtual and Physical Prototyping,* ***18***(1), p. e2210553.

[9] Sato, M. M., Wong, V. W. H., Law, K. H., Yeung, H., Yang, Z., Lane, B., and Witherell, P., 2022. "Anomaly detection of laser powder bed fusion melt pool images using combined unsupervised and supervised learning methods". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 86212, American Society of Mechanical Engineers, p. V002T02A070.

[10] Surovi, N. A., Hussain, S., and Soh, G. S., 2022. "A study of machine learning framework for enabling early defect detection in wire arc additive manufacturing processes". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 86229, American Society of Mechanical Engineers, p. V03AT03A002.

[11] Samie Tootooni, M., Dsouza, A., Donovan, R., Rao, P. K., Kong, Z., and Borgesen, P., 2017. "Classifying the dimensional variation in additive manufactured parts from laser-scanned three-dimensional point cloud data using machine learning approaches". *Journal of Manufacturing Science and Engineering,* ***139***(9), p. 091005.

[12] Liu, J., Liu, C., Bai, Y., Rao, P., Williams, C. B., and Kong, Z., 2019. "Layer-wise spatial modeling of porosity in additive manufacturing". *IISE Transactions,* ***51***(2), pp. 109–123.

[13] Surovi, N. A., and Soh, G. S., 2023. "A heuristic approach to classify geometrically defective bead segments based on range of curvature, range of sound power and maximum height". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 87301, American Society of Mechanical Engineers, p. V03AT03A005.

[14] Petrich, J., Snow, Z., Corbin, D., and Reutzel, E. W., 2021. "Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing". *Additive Manufacturing,* ***48***, p. 102364.

[15] Razvi, S. S., Feng, S., Narayanan, A., Lee, Y.-T. T., and Witherell, P., 2019. "A review of machine learning applications in additive manufacturing". In International design engineering technical conferences and computers and information in engineering conference, Vol. 59179, American Society of Mechanical Engineers, p. V001T02A040.

[16] Sakirin, T., and Kusuma, S., 2023. "A survey of generative artificial intelligence techniques". *Babylonian Journal of Artificial Intelligence,* ***2023***, pp. 10–14.

[17] Bandi, A., Adapa, P. V. S. R., and Kuchi, Y. E. V. P. K., 2023. "The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges". *Future Internet,* ***15***(8), p. 260.

[18] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., and Sun, L., 2023. "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt". *arXiv preprint arXiv:2303.04226*.

[19] McClelland, R., 2022. "Generative design and digital manufacturing: Using ai and robots to build lightweight instruments". In SPIE Optics and Photonics.

[20] Elbadawi, M., Li, H., Sun, S., Alkahtani, M. E., Basit, A. W., and Gaisford, S., 2024. "Artificial intelligence generates novel 3d printing formulations". *Applied Materials Today,* ***36***, p. 102061.

[21] Hertlein, N., Buskohl, P. R., Gillman, A., Vemaganti, K., and Anand, S., 2021. "Generative adversarial network for early-stage design flexibility in topology optimization for additive manufacturing". *Journal of Manufacturing Systems,* ***59***, pp. 675–685.

[22] Petrik, J., Kavas, B., and Bambach, M., 2023. "Meltpoolgan: Auxiliary classifier generative adversarial network for melt pool classification and generation of laser power, scan speed and scan direction in laser powder bed fusion". *Additive Manufacturing,* ***78***, p. 103868.

[23] Mu, H., He, F., Yuan, L., Hatamian, H., Commins, P., and Pan, Z., 2024. "Online distortion simulation using generative machine learning models: A step toward digital twin of metallic additive manufacturing". *Journal of Industrial Information Integration*, p. 100563.

[24] Liu, W., Wang, Z., Tian, L., Lauria, S., and Liu, X., 2021. "Melt pool segmentation for additive manufacturing: A generative adversarial network approach". *Computers & Electrical Engineering,* ***92***, p. 107183.

[25] Badini, S., Regondi, S., Frontoni, E., and Pugliese, R., 2023. "Assessing the capabilities of chatgpt to improve additive manufacturing troubleshooting". *Advanced Industrial and Engineering Polymer Research,* ***6***(3), pp. 278–287.

[26] Jignasu, A., Marshall, K., Ganapathysubramanian, B., Balu, A., Hegde, C., and Krishnamurthy, A., 2023. "Towards foundational ai models for additive manufacturing: Language models for g-code debugging, manipulation, and comprehension". *arXiv preprint arXiv:2309.02465*.

[27] Fang, Y., Chen, M., Liang, W., Zhou, Z., and Liu, X., 2023. "Knowledge graph learning for vehicle additive manufacturing of recycled metal powder". *World Electric Vehicle Journal,* ***14***(10), p. 289.

[28] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774*.

[29] OpenAI, 2023. Gpt-4v(ision) system card.

[30] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. "Language models are few-shot learners". *Advances in neural information processing systems,* ***33***, pp. 1877–1901.

[31] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al., 2022. "Lamda: Language models for dialog applications". *arXiv preprint arXiv:2201.08239*.

[32] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. "Llama 2: Open foundation and fine-tuned chat models". *arXiv preprint arXiv:2307.09288*.

[33] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I., 2021. "Zero-shot text-to-image generation". In International Conference on Machine Learning, PMLR, pp. 8821–8831.

[34] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., 2022. "High-resolution image synthesis with latent diffusion models". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

[35] Kim, D. B., Witherell, P., Lipman, R., and Feng, S. C., 2015. "Streamlining the additive manufacturing digital spectrum: A systems approach". *Additive manufacturing,* ***5***, pp. 20–30.

[36] Yeung, H., Yang, Z., and Yan, L., 2020. "A meltpool prediction based scan strategy for powder bed fusion additive manufacturing". *Additive Manufacturing,* ***35***, p. 101383.

[37] Yang, Z., Lu, Y., Yeung, H., and Krishnamurty, S., 2019. "Investigation of deep learning for real-time melt pool classification in additive manufacturing". In 2019 IEEE 15th international conference on automation science and engineering (CASE), IEEE, pp. 640–647.

[38] Zhan, Z., and Li, H., 2021. "Machine learning based fatigue life prediction with effects of additive manufacturing process parameters for printed ss

316l". *International Journal of Fatigue,* ***142***, p. 105941.

[39] Surovi, N. A., and Soh, G. S., 2022. "Process map generation of geometrically uniform beads using support vector machine". *Materials Today: Proceedings,* ***70***, pp. 113–118.

[40] AbouelNour, Y., and Gupta, N., 2022. "In-situ monitoring of sub-surface and internal defects in additive manufacturing: A review". *Materials & Design*, p. 111063.

[41] Surovi, N., and Soh, G. "Multi-bead and multi-layer printing geometric defect identification using single bead trained models".

[42] Surovi, N. A., Dharmawan, A. G., and Soh, G. S., 2021. "A study on the acoustic signal based frameworks for the real-time identification of geometrically defective wire arc bead". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 85383, American Society of Mechanical Engineers, p. V03AT03A003.

[43] Bartlett, J. L., Jarama, A., Jones, J., and Li, X., 2020. "Prediction of microstructural defects in additive manufacturing from powder bed quality using digital image correlation". *Materials Science and Engineering: A,* ***794***, p. 140002.

[44] Huang, J., Kwok, T.-H., Zhou, C., and Xu, W., 2019. "Surfel convolutional neural network for support detection in additive manufacturing". *The International Journal of Advanced Manufacturing Technology,* ***105***, pp. 3593–3604.