

2-Exploring AI Image Generators' Deep Understanding of English Language Prompts

استكشاف الفهم العميق لمولدات صور الذكاء الاصطناعي للموجهات النصية في اللغة الإنجليزية



Dr. Fatemah Bazzi

الدكتورة فاطمة بزي

An Associate Professor of English Language and Literature at the Lebanese University, Faculty of Literature and Human Sciences, First Branch

أستاذ مساعد في اللغة الإنكليزية وآدابها في الجامعة اللبنانية،

كلية الآداب والعلوم الإنسانية.

ms.bazzi_12@gmail.com

Abstract:

This research was done to explore whether Artificial Intelligence (AI) has a deep level of language understanding. Lately, many researches have been done to examine how AI systems generate images when provided with descriptive prompts or texts.

These researches have focused on the comparison between AI image generators to check the quality of the images produced and to what extent they are close to reality. Nevertheless, investigating the alignment between the written input and the images generated and evaluating the capability of image generators to deeply understand English language have been limited. Consequently, the researcher of the current study conducted a text-to-image qualitative and quantitative analyses of two popular AI image generators, DALL-E and Gencraft, to explore the semantic alignment between the prompts provided and the images generated. This was done in an attempt to evaluate the capability of AI image generators to deeply understand English language. The results revealed that the content of the images generated did not semantically align with the input provided, mainly when inferring contextual details and understanding complete sentences vs. phrases.

Keywords: artificial intelligence, text-to-image generator, prompts, semantic alignment

نبذة عن البحث باللغة العربية:

تم إجراء هذا البحث لمعرفة ما إذا كان الذكاء الاصطناعي يمتلك القدرة على الفهم العميق للغة. في الآونة الأخيرة، تم إجراء عدد من الأبحاث لمعرفة كيفية عمل أنظمة الذكاء الاصطناعي على إنشاء الصور عندما يتم تزويدها بموجهات أو نصوص وصفية. قامت هذه الأبحاث بالتركيز على المقارنة بين مولدات الصور المعتمدة على الذكاء الاصطناعي للتأكد من جودة الصور المنتجة والى أي حد هي قريبة من الواقع. وبالرغم من ذلك، فإن التحقق من وجود توافق بين المدخلات المكتوبة والصور التي تم إنشاؤها وتقييم قدرة مولدات الصور على فهم اللغة الإنجليزية بشكل عميق لا يزال محدوداً. لذلك قامت الباحثة من خلال هذه الدراسة بإجراء تحليل نوعي وكَمّي حول

كيفية تحويل النص الى صورة من خلال استخدام مولّدي صور الذكاء الاصطناعي Gencraft و DALL-E وهما شائعا الاستخدام وذلك بهدف الكشف عن مدى وجود محاذاة دلالية بين الموجهات التي يتم تزويد مولدات الصور بها وبين الصور المنتجة. والهدف من هذا التحليل هو تقييم قدرة مولدات صور الذكاء الاصطناعي على فهم اللغة الإنجليزية بشكل عميق. وقد أظهرت نتائج هذا البحث أنه لم يكن هناك محاذاة لغوية بين محتوى الصور والمدخلات المستخدمة بالتحديد عند استنتاج التفاصيل السياقية وفهم الجمل الكاملة مقابل العبارات.

الكلمات المفتاحية: الذكاء الاصطناعي، مولد صور من النص، مدخلات نصية، محاذاة دلالية.

1. Introduction

In the past few years, a new era of using AI in many domains has started. A specific field of “large language” artificial intelligence algorithm has been elaborated with a high level of fluency of using the English language (Warner, 2022). This has led to many debates on whether AI systems can really understand grammar, sentence structure, and every word of the language to the extent that some researches have questioned the capability of AI to comprehend the language like humans. Some AI systems have been used to provide answers for certain questions or language tasks in the field of education, so they have been programmed with the rules of English grammar and syntax. Other AI systems have been developed to be used in other fields, such as business, medicine, art, among many others. Specifically, in the field of art, AI image generators have been used to produce images based on the written prompts or texts provided to the system. As for the efforts to use machines for generating images from written texts, it can be traced to the

times of deep generative models introduced by Mansimov et al. (2016) when they added text information to DRAW, which is a recurrent neural network for image generation (Gregor et al. 2015). After that, many AI image generators have been developed to dominate this task.

However, the capability of these image generators to fully understand the human language has been subjected to criticisms. Precisely, the debates have revolved around whether image generators can provide answers to certain language prompts without having a deep understanding of semantics (Levesque et al., 2012) or having features that require higher levels of cognitive skills (Bonnici et al., 2016). In other words, AI and computers can do amazing tasks; however, according to many researchers, they are still far from being able to understand what humans say. Thus, they require both a powerful generative model and cross-modal understanding (Ding et al., 2021). Consequently, examining text-to-image generation is a good topic for language researchers to investigate the issues or limitations of AI image generators in this domain. That is why the main purpose of the current study is to investigate the deep level of understanding English language prompts by AI image generators. This investigation is based on identifying AI challenging capabilities of understanding the language by referring to two popular image generators, DALL-E and Gencraft, to explore whether they can develop semantic alignment between the prompts provided and the images generated.

2. Literature Review

The huge and rapid evolution in the field of AI has paved the way for many text-to-image generators to develop, by providing users with different innovative features. Many studies have examined the power and limitations of these AI image generators in an attempt to evaluate the quality of their output. However, before focusing on these studies, it is essential to examine the approaches related to language processing by AI systems. According to Jia et al. (2021), the revision of the existing literature has shown that the study of visual and vision-language representation learning has been commonly done separately using diverse training of data sources. On one hand, Jia et al. explained, the vision domain requires pre-training on large-scale supervised data and heavy work on data gathering, sampling, and human annotation; thus, it is very challenging to be scaled. On the other hand, vision-language pre-training datasets necessitate more work and effort on human annotation, semantic parsing, cleaning, and balancing. Therefore, it is significant to know how language processing of AI systems functions in an attempt to examine the capabilities, challenges, and limitations that have been identified.

Recently, more advanced approaches and methods of language processing by AI systems have been proposed, such as the context of contemporary Natural Language Processing (NLP). A survey on NLP history done by Brock (2018) showed two main approaches of research, recognition and reasoning. Under the first approach, Brock mentioned deep learning and making

judgment, which is based on the count of character combinations such as words. As for the reasoning approach, it is based on Natural Language Understanding (NLU), which aims for deep understanding of human cognition and tries to deduce by using processing knowledge such as syntax, semantics, and common sense (Micelli et al., 2009). Hence, the field of NLP has taken a huge progress to an extent that AI systems can generate convincing passages with the push of a button. Nevertheless, researches and tests have been conducted to assess to what extent they can understand language. Levesque et al., (2012) developed the test of Winograd Schema Challenge to evaluate the common-sense reasoning of NLP systems. The main aim of their test was to find out whether certain language problems could be answered without a deeper grasp of semantics. They found that some state-of-the-art deep-learning models can reach around 90% accuracy. However, with other language problems, the performance fell between 59.4% and 79.1%; by contrast, humans still reached 94% accuracy. Their findings proved that to have more common sense, language processing systems should include other techniques, such as structured knowledge models (Hao, 2020). According to Zhang et al. (2018), statistical NLP methods are usually trained on large data, including millions of word occurrences, so they can process considerably large inputs.

However, the problem with these methods is that they have limited accuracy when they are provided with challenging tasks involving deep semantics or common sense reasoning (Richard-Bollans et al., 2018). Another problem identified with statis-

tical NLP is its limited transparency at certain cases, for it is often difficult to identify the exact ideas that led an AI system to take certain choices (Brock, 2018). It is worth adding that NLU by AI systems requires two main processes, data pre-processing and algorithm development. As for data pre-processing, it includes four main steps: breaking down the sentences or text into tiny units, removing common words and keeping the unique ones, simplifying words to their root forms, and tagging words based on parts of speech like nouns, verbs, and adjectives (Nayak et al., 2016). That is why the current study is based on breaking down prompts into tiny units to examine the semantic alignment between the textual prompts and the images generated.

Moving to the field of image synthesis and the use of AI image generators, getting quality illustrative images requires specific skills so that users can write the accurate text prompt. Recently, generative models have gained the ability to generate human-like natural language (Brown et al., 2020), infinite high-quality synthetic images (Karras et al., 2020) and highly diverse human speech and music (Dhariwal et al., 2020). These models can be used in many different ways, such as generating images from text prompts or learning useful feature representations (Donahue & Simonyan, 2019). However, although these models are able to produce realistic images and sounds, a lot of improvement is still required beyond the present state-of-the-art (Dhariwal & Nichol, 2021). In their study, Crowson et al. (2022) applied a new method of generating and manipulating images that are based on prompts written by humans, and they concluded that when the textual prompt and image content have

low semantic similarity, the quality of visual images is higher. According to Ding et al. (2021), generating images from texts requires many important things from the system: differentiating between shapes, colors, gestures, and other features from pixels; comprehending the input text; aligning items and features with their matching words and their synonyms; and knowing how to deal with complex distributions in an attempt to produce the overlapping and combined items and features that are beyond basic visual functions (Grill-Spector & Malach, 2004), which requires higher levels of cognitive skills (Bonnici et al., 2016).

Based on the previously mentioned findings, the major developments in text-to-image generators may have led many people and experts to think that AI image generators have the capability to do anything, but this is not the case (Lloyd, 2019). As such, examining whether they have deep understanding and can generate images that accurately align with the provided prompts is very significant. That is why the current study explores the capability of two AI image generators to deeply understand English language. The two generators were chosen because they depend mainly on transforming English prompts into images, and they have millions of users. DALL-E is a popular AI image generator, with over 1.5M users producing more than 2M images per day. This generator is trained to produce images from text captions for a wide variety of concepts that are expressed in natural language. The main difference between it and the traditional image synthesis techniques is that it takes advantage of the extensive data it has been trained on to produce new images that never existed before (Noor, 2023). Unlike Gencraft,

it does not allow users to type their own prompts. Instead, it allows users to change certain words of existed prompts for each category available. Actually, it works by receiving both the text and the image as a single stream of 1280 tokens—256 for the text and 1024 for the image—and models all of them autoregressively. A token is any symbol from a discrete vocabulary, and DALL-E’s vocabulary includes tokens for texts and image concepts (OpenAI, 2021).

As for Gencraft, it is an AI Art Generator that transforms descriptive prompts to vibrant output, images and videos, within 15–20 seconds. Gencraft allows millions of users to generate personalized art photos from a few words, and it is available across many devices, such as web, iOS, and Android platforms. Moreover, Gencraft allows users to type their own prompts and offers many styles that can be combined with text prompts to increase the creativity while generating images. Its users can write the description of the image up to 250 characters, with an additional 250 characters for extra details if necessary (Asad, 2023). It is worth adding that this study is different from those studies that have compared AI image generators to evaluate the quality of the generated images. The aim of the current study is to evaluate to what extent such DALL-E and Gencraft have a deep understanding of the language by examining their capability to generate images that semantically align with the textual prompts provided. In addition, in order to deviate from previous studies done on AI language understanding, the researcher chose two capabilities to examine: inferring contextual details and understanding complete sentences vs. phrases by DALL-E and Gencraft.

3. Research Questions

The current study explores two research questions:

(Q.1) To what extent do AI image generators have the capability to infer the contextual details of English language prompts?

(Q.2) To what extent do AI image generators have the capability to understand complete sentences vs. phrases of English language prompts?

4. Methodology, Participants, and Procedure

The purpose of this research is to investigate whether AI image generators have a deep level of language understanding. The research follows the mixed-method approach that examines to what extent two popular AI image generators, Gencraft and DALL-E, are capable to produce images that semantically align with the textual prompts provided in English language. Qualitative and quantitative data were collected to help answer the research questions. First, a set of written prompts was used by the researcher to examine two capabilities: inferring contextual details and understanding complete sentences vs. phrases. To explore the first capability, the researcher used the features of DALL-E for providing prompts, based on its available categories. The same content was used to write descriptive prompts on Gencraft since it allows users to write the whole prompts instead of choosing concepts from given lists.

According to the second capability, the researcher examined it by referring to Gencraft because DALL-E does not have the options of writing complete sentences by users. The researcher

wrote the same content in two different structures, complete sentences vs. phrases, and with two different options, cartoon and real images. That is because she wanted to check more than one style of image production. In addition, the researcher's aim was to use a specific content, so she included specific details, such as numbers, compound nouns, shapes, and time. Besides, two different tenses were used for writing the complete sentences, present tense and past tense, so that not to limit the content to one tense.

Moreover, to cross validate the results of the qualitative data, the researcher randomly selected 50 images from each generator, based on different users' prompts. The images and their corresponding prompts were sent in a form of questionnaire to 10 people, who commonly use different image generators. They had to evaluate the semantic alignment between each image content and its prompt. After that, a qualitative and quantitative analyses of the data collected took place, where the researcher studied the semantic similarity between each word of the prompts and the content of the images generated. In addition, she analyzed the answers of the questionnaire using Excel Sheets. Finally, tables and figures were used to provide the results.

5. Data Collection and Analysis

Data analysis included many steps. First, a comparison between each input and its output was done to evaluate the semantic alignment between the two in order to draw out conclusions on the capability of AI image generators to deeply understand the

language. This analysis was done in terms of two main parts: a qualitative analysis with respect to AI image generators' capability to infer contextual details and a qualitative analysis with respect to AI image generators' understanding of complete sentences vs. phrases. As for the quantitative analysis, it focused on the analysis of the questionnaire with respect to the semantic alignment between image content and textual prompts. Finally, the qualitative data were presented in tables and anecdotal analysis, where each one of the two capabilities was analyzed separately to help the researcher provide answers for the research questions. Regarding the quantitative data, they were presented in figures and statistical analysis.


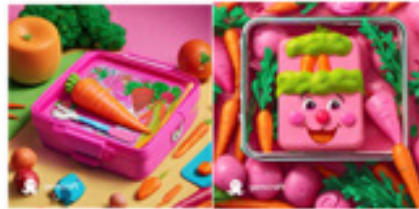

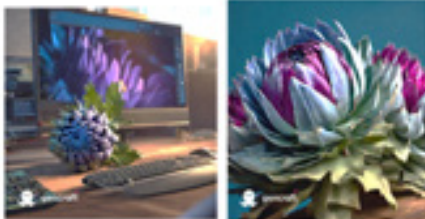
5.1. Qualitative analysis with respect to AI image generators' capability to infer contextual details

The first part of data analysis explored the capability of AI image generators to infer contextual details when they are provided with prompts with precise details or precise situations. Based on the possibilities and tokens provided by DALL-E, the researcher chose two categories: an image of an object that requires inference (specific style of color, shape, number, etc.) and an image of an object with specific text written on it. In the first category, two prompts were provided: "a plastic lunchbox with an image of a pink carrot" and "a computer monitor with an image of a blue artichoke". In the second category, also two prompts were provided: "t-shirt that has the word 'meatloaf' written on it" and "grocery bag with the word 'peekaboo' written on it". The same prompts were written on the generator of Gencraft.

As table 1 reveals, DALL-E provided many images for each prompt, while Gencraft provided two images for each one. The first prompt included five specific details that should be shown in the output: (plastic) + (lunchbox) + (image) + (pink) + (carrot). Concerning the output provided by DALL-E, some of the images included content that did not semantically match with the details provided: an orange or violet carrot instead of a pink one, a real carrot instead of its image, and the absence of a lunchbox. As for the results provided by Gencraft, the content of the two images did not semantically match with certain details provided: real carrots and images of carrots, pink and orange carrots, and the presence of other vegetables. In addition, one of the images did not accurately show the lunchbox.

As for the second prompt, it included five specific details that should be shown in the output: (a computer) + (monitor) + (image) + (blue) + (an artichoke). DALL-E provided many images, some of which included content that did not semantically match with the details provided: many artichokes instead of one, a real artichoke instead of its image, and a green artichoke instead of a blue one. As for the results provided by Gencraft, one image included real artichoke and did not include a computer monitor. Besides, the other image showed a real artichoke with its image reflected by the computer monitor, but their color was violet instead of being blue. Hence, although some details were accurately revealed in the output, the content of the images generated by both AI did not semantically align with inferential details of the textual prompts.





Table 1: Images produced by AI image generators for objects with inferential details

<i>DALL-E</i>	<i>Gencraft</i>
	
	

As for the second category, the first prompt included four precise details that should be shown in the generated images: (t-shirt) + (word) + (meatloaf) + (written). As revealed by table 2, the images provided by DALL-E included content that did not semantically match with the details provided. That is because the word “meatloaf” was misspelled in many images: “meant”, “mealeof”, “meatlaf”, “meatlep”, and “meatlie”. As for the results provided by Gencraft, the two images also included misspelled words: “metloaf” and another word that had two letters merged together, “t” and “l”. In addition, one of the images did not include t-shirt with the word written on it; instead, it revealed a bowl of meat with the word under it.

In addition, the second prompt included five precise details: (grocery) + (bag) + (word) + (peekaboo) + (written). Similarly, the images provided by DALL-E included the word “peekaboo” misspelled many times: “peekoboo”, “peekboo”, “peegkaboo”, “peekago”, and “peekao”. Likewise, the two images provided by Gencraft revealed misspelled words: “peeb’oo” and “peekebo”. Therefore, the capability of generating images of an object with specific text written on it resulted in spelling errors by both AI image generators. This proved that the content of the images produced by both AI generators did not semantically align with all the details of the textual prompts provided, the fact that reveals limited understanding of the language used.

Table 2: Images produced by AI image generators for objects with specific written text

<i>DALL-E</i>	<i>Gencraft</i>
	
	





5.2. Qualitative analysis with respect to AI image generators’ understanding of complete sentences vs. phrases

The second part of data analysis examined the capability of AI image generators to understand the content of complete sentences vs. phrases. Four pairs of prompts were used. The complete sentences of the first two pairs had verbs in the simple present tense, while those of the other two pairs had verbs in the past tense. The first pair of prompts included “A little boy holds three bottles” vs. “A little boy holding three bottles”. The content of the two prompts included five specific details: (little) + (boy) + (holds/holding) + (three) + (bottles). As shown by table 3, two images were provided for each prompt. The two images provided for both the complete sentence and the phrase included three details that semantically matched with the input: (little) + (boy) + (bottles). However, the images provided for the complete sentence did not accurately show the other details: the number and the act of holding three bottles. That is because in both images the little boy was surrounded by many bottles, held one bottle in the first image, and held two bottles in the second. On the other hand, the number was shown more accurately in the images provided for the phrases, three bottles in the first image and four bottles in the second one. However, in both images, the little boy held one bottle.

As for the second pair of prompts, it included “An old man walks with four dogs” vs. “An old man walking with four dogs”. The content of both prompts included five specific details: (old) + (man) + (walks/walking) + (four) + (dogs). As shown by table 3, the two images provided for both the complete sentence and the phrase included three details that semantically aligned with the prompts: (old) + (man) + (walks/walking) + (dogs). On one

hand, one of the images provided for the complete sentence transferred the number of dogs into the image accurately; however, the second image included five dogs instead of four. On the other hand, the two images provided for the phrase included five and six dogs, respectively. As such, Gencraft showed better understanding of phrases vs. complete sentences in the first pair of prompts. As for the second one, the generator did not show better understanding of the content of phrases vs. that of complete sentences. However, the numbers in the provided input did not semantically align with those shown in the output.

Table 3: Images produced by AI image generators for complete sentences in the present tense vs. phrases

<i>Complete Sentence</i>	<i>Phrase</i>
A little boy holds three bottles.	A little boy holding three bottles
	
An old man walks with four dogs.	An old man walking with four dogs
	



The third pair of prompts included “Two maids used the washing machine at night” vs. “Two maids using the washing machine at night”. The content of the two prompts included five specific details: (two) + (maids) + (used/using) + (washing machine) +

(night). As revealed by table 4, the two images provided for both the complete sentence and the phrase included three details that semantically matched with the input: (maid) + (used/using) + (washing machine). However, the images provided for the complete sentence did not accurately show the other details: the number and the time required, which is “night”. In addition, one image showed two washing machines instead of one. On the other hand, the images provided for the phrases revealed the accurate number, two maids. Nevertheless, in one of these images, one maid was shown using the washing machine, while the other one was inside it. As for the temporal knowledge provided, it was inaccurately transformed for the complete sentence because one of its images revealed day time. On the other hand, one of the images for the phrase showed night time.

As for the last pair of prompts, it included “A fisherman ate seafood in an oval plate” vs. “A fisherman eating seafood in an oval plate”. The content of both prompts included five specific details: (fisherman) + (ate/eating) + (seafood) + (oval) + (plate). As shown by table 4, the two images provided for both the complete sentence and the phrase included two details that semantically aligned with the prompts: (seafood) + (plate). On one hand, the two images provided for the complete sentence did not semantically align with the three other details: fisherman, the action of eating, and the oval shape. On the other hand, one of the images for the phrase showed a fisherman, but he was not eating. As for the second image for the phrase, it revealed only the hands of a man. Regarding the oval shape, it was not revealed by any image. As such, Gencraft showed

more semantic similarity between the content of phrases and their generated images than between complete sentences and their generated images. In addition, it did not transfer the input related to shape and the verb of eating into the content of the images provided for both sentences and phrases.

Table 4: Images produced by AI image generators for complete sentences in the past tense vs. phrases

<i>Complete Sentence</i>	<i>Phrase</i>
Two maids used the washing machine at night.	Two maids using the washing machine at night
	
A fisherman ate seafood in an oval plate.	A fisherman eating seafood in an oval plate
	

5.3. Quantitative analysis of the questionnaire with respect to the semantic alignment between image content and textual prompts by AI image generators

The third part of data analysis examined the images found on the two generators, DALL-E and Gencraft, to evaluate the semantic alignment of the textual prompts and the output provided by the generators. A questionnaire of two parts was sent to 10 users of different image generators. Each part comprised 50 images

with their prompts from DALL-E and Gencraft to be evaluated by the users. They had to evaluate the content of each image by responding to one question: Does the image accurately reveal what its textual prompt states? Then they had to choose an answer from a Three-point Likert Scale: “yes”, “somewhat”, or “no”. The results of DALL-E were revealed in figure 1. Based on the results of the first part of the questionnaire, the number of images with content that did not semantically align with the prompts ranged between 6 and 19 out of 50, and their average number was 12.1 images (24.2%). As for the number of images with content that “somewhat” semantically aligned with the prompts, it ranged between 21 and 39 out of 50, and their average number was 31.8 images (63.6%). Finally, the number of images with content that semantically aligned with the prompts ranged between 0 and 10 out of 50, and their average number was 6.1 images (12.2%). Consequently, the majority of users considered the majority of DALL-E images (63.6%) “somewhat” semantically aligned with the textual prompts provided.

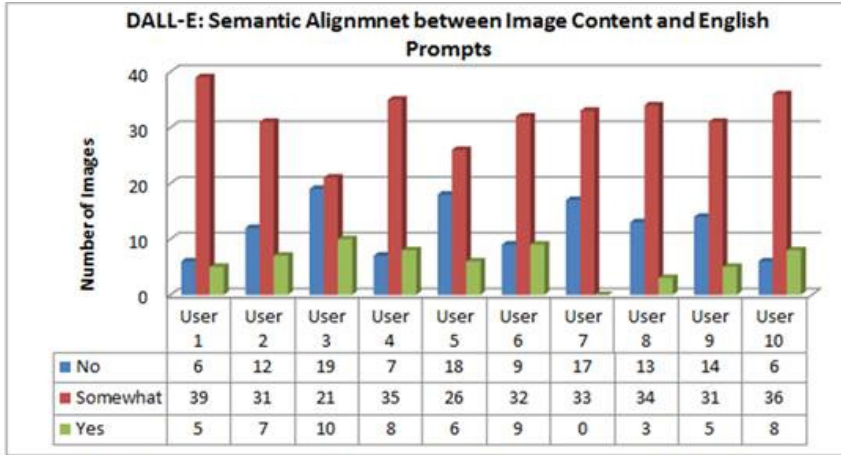


Figure 1. Results of questionnaire on the semantic alignment between DALL-E images and their prompts

As for the results of Gencraft, they are revealed in figure 2. Based on the results of the second part of the questionnaire, the number of images with content that did not semantically align with the prompts ranged between 1 and 9 out of 50, and their average number was 5.5 images (11%). As for the number of images with content that “somewhat” semantically aligned with the prompts, it ranged between 22 and 36 out of 50, and their average number was 30.7 images (61.4%). Finally, the number of images with content that semantically aligned with the prompts ranged between 9 and 19 out of 50, and their average number was 13.8 images (27.6%). Accordingly, the majority of users considered the majority of Gencraft images (61.4%) “somewhat” semantically aligned with the textual prompts provided.

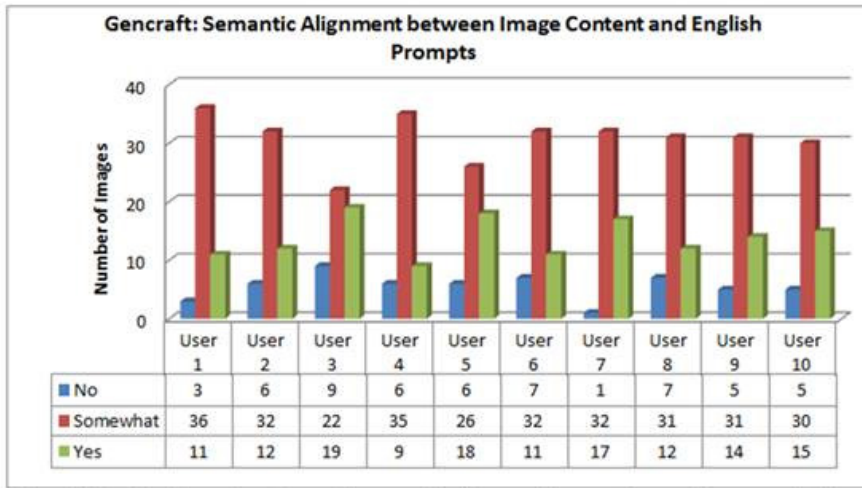


Figure 2. Results of questionnaire on the semantic alignment between Gencraft images and their prompts

Finally, figure 3 sumps up the statistical analysis of the two parts of the questionnaire. Very close percentages represented the capability of both generators to provide a “somewhat” semantic alignment between image content and textual prompts: 63.6% by DALL-E and 61.4% by Gencraft. However, the statistical analysis showed a difference with respect to “no” semantic alignment between image content and textual prompts: 24.2% by DALL-E while 11% by Gencraft. Likewise, different results were provided with respect to “yes” semantic alignment between image content and textual prompts: 12.2% by DALL-E while 27.6% by Gencraft. Therefore, although the semantic alignment between image content and textual prompts were more revealed by Gencraft images, the majority of images provided by both generators did not show a complete semantic alignment between image content and English prompts.

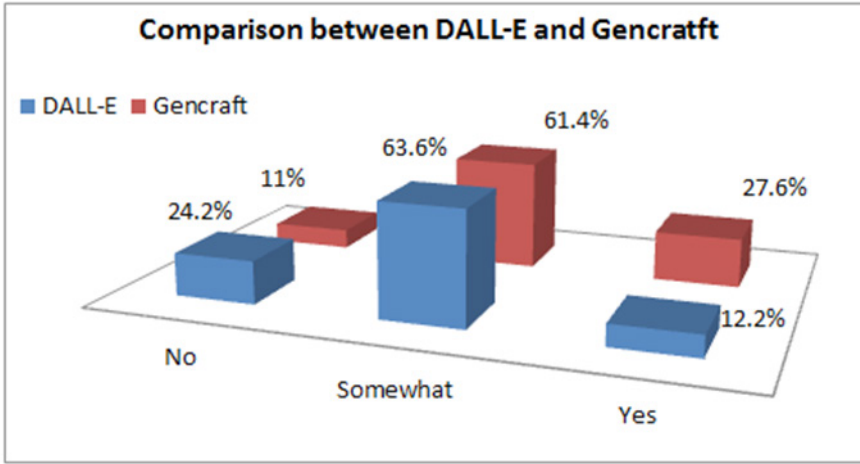


Figure 3. Comparison between DALL–E and Gencraft with respect to the semantic alignment between images and their prompts

6. Results and Discussion

The present study examined two main questions. The first one investigated the extent to which AI image generators have the capability to infer the contextual details of English language prompts. As for the second one, it studied the extent to which AI image generators have the capability to understand complete sentences vs. phrases of English language prompts. The results of the qualitative and quantitative data analyses helped the researcher to provide answers to the two questions by using two popular AI image generators, DALL–E and Gencraft.

According to research Q.1, when provided with prompts with inferential contextual details, both AI image generators produced images that did not semantically align with specific details. That was revealed by the qualitative and quantitative data analyses. First, different colors were provided, and some words were not

revealed by some images. Second, some words appeared with spelling errors in the images generated. In addition, the location and nature of some details did not semantically match with what the prompts had provided. Referring to the results of Crowson et al. (2022) that the quality of visual images is higher when there is low semantic similarity between the textual prompt, the researcher concluded that image generators focus more on the quality of the images generated. The lack of semantic alignment in some of the images generated justified the previous result. Moreover, the results compiled with the findings of Ding et al. (2021) and Bonnici et al. (2016) that the ability to differentiate between shapes, colors, gestures, and other features and to align items and features with their matching words are matters that necessitate higher levels of cognitive skills. Consequently, AI image generators have limited levels of the cognitive skills required to accomplish the task of aligning between the images and all the contextual details provided.

As for research Q2., qualitative data analysis of the prompts used to evaluate the understanding of complete sentences vs. phrases proved that AI image generators did not accurately comprehend complete sentences. Many specific details in the content of the images did not semantically match with the textual prompts: numbers, shapes, time, and certain action verbs, mainly in the past tense. On the other hand, better understanding of phrases was evident by the image generator. This agreed with the findings of Dhariwal and Nichol (2021) that despite the fact that image generators can produce realistic images and sounds, many improvements are still needed in the field of AI art production. Therefore, AI image generators do not have the

capability to grasp the meaning provided by complete sentences. As for the quantitative data analysis, it was evident that AI image generators cannot semantically transform all the details provided into images with the required content. These results concurred with Richard–Bollans et al.’s findings (2018) regarding the main problem of NLP methods, which reveals limited accuracy when provided with challenging tasks that include deep semantics or common sense reasoning.

7. Conclusion

The current study verified that AI image generators, such as DALL–E and Gencraft, have limited cognitive skills when they transform the content of English language prompts into images. Evaluating the semantic alignment between the written input and the image content resulted in showing the limited capability of AI image generators to deeply understand the language, specifically when provided with contextual details that require inference. In addition, the capability of the image generators to comprehend complete sentences vs. phrases with similar content is also limited mainly in terms of transferring numbers, shapes, and temporal knowledge into images. Therefore, with the plenty of development and progress done in the field of AI, having a deep level of language understanding is still a challenging task for AI image generators.

References

- Asad, M. (2023). Gencraft – A comprehensive analysis of the text-to-image generation AI tool. Divine.AI. <https://divine.ai/blog/gencraft-a-comprehensive-analysis-of-the-text-to-image-generation-ai-tool/>
- Bonnici, H., Richter, F., Yazar, Y., & Simons, S. (2016). Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *Journal of Neuroscience*, 36(20), 5462–5471.
- Brock, D. C. (2018). Learning from artificial intelligence’s previous awakenings: The history of expert systems. *AI Magazine*, 39(3), 3–15.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castriato, L., & Raff, E. (2022). Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision* (pp. 88–105). Cham: Springer Nature Switzerland.
- Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.
- Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794.
- Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., ... & Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 19822–19835.

- Donahue, J., & Simonyan, K. (2019). Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*, 32.
- Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning* (pp. 1462–1471). PMLR.
- Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27, 649–677.
- Hao, K. (2020). AI still doesn't have the common sense to understand human language. *MIT Technology Review*.
- Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* (pp. 4904–4916). PMLR.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110–8119).
- Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Lloyd, S. (2019). Wrong, but more relevant than ever. In Brockman, J. (Ed.), *Possible Minds: Twenty-Five Ways of Looking at AI* (1–12). New York: Penguin Press.
- Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*.
- Micelli, V., van Trijp, R., and De Beule, J. (2009). Framing fluid construction grammar. In *The 31th Annual Conference of the Cognitive Science Society* (pp. 3023–3027).
- Nayak, A. S., Kanive, A. P., Chandavekar, N., & Balasubramani, R. (2016). Survey on pre-processing techniques for text mining. *International Journal of Engineering and Computer Science*, 5(6), 16875–16879.
- Noor, K. (2023). Dall-e statistics: The power of artificial imagination. *Market Splash*. <https://marketsplash.com/>
- OpenAI, (2021). DALL·E: Creating images from text. <https://openai.com/research/dall-e>

Richard–Bollans, A., Gomez Alvarez, L., and Cohn, A. G. (2018). The role of pragmatics in solving the winograd schema challenge. In Proceedings of the Thirteenth International Symposium on Commonsense Reasoning. CEUR Workshop Proceedings.

Warner, J. (2022). Language AI don't know no grammar. Inside Higher Ed. <https://www.insidehighered.com/opinion/blogs>

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1253