UNIVERSITY OF TARTU

Institute of Computer Science

Computer Science Curriculum

Helena Lindström

Systematic Analysis on GPT-4o, DALL·E 3 and Stable Diffusion 3.5. Do Text-to-image Models Incorporate World Models?

Bachelor's Thesis (9 ECTS)

Supervisor: Jaan Aru, PhD

Tartu 2025

Systematic Analysis on GPT-4o, DALL·E 3 and Stable Diffusion 3.5. Do Text-to-image Models Incorporate World Models?

Abstract:

Text-to-image models intend to create correct images based on given prompts. This research aims to determine whether studied text-to-image models have integrated world models. Text-to-image models GPT-4o, DALL·E 3 and Stable Diffusion 3.5 generate images based on prompts. For this study, 25 objects that follow a specific logic have been selected, and based on these objects, 25 prompts have been constructed. Based on this study, GPT-4o appeared to be the best in depicting selected objects as 31% of the generated images were correct. Only 12% of images created by DALL·E 3 and 11% of images created by Stable Diffusion 3.5 were correct. In conclusion, due to poor results, it can be stated that GPT-4o, DALL·E 3 and Stable Diffusion 3.5 have not incorporated world models.

Keywords:

Artificial intelligence, text-to-image model, world models

CERCS: P176 Artificial intelligence

Süstemaatiline analüüs GPT-4o, DALL·E 3 ja Stable Diffusion 3.5 põhjal. Kas pildigeneraatorid integreerivad mudeleid maailma kohta?

Lühikokkuvõte:

Pildigeneraatorite peamine eesmärk on etteantud lähtetekstide põhjal korrektseid pilte luua. Käesolevas töös uuritakse, kas valitud pildigeneraatorid (GPT-4o, DALL·E 3 ja Stable Diffusion 3.5) põhinevad piltide loomisel maailma mudelitele. Töö jaoks on valitud 25 objekti, millest igaüks järgib kindlaid toimimispõhimõtteid, ja objektidest piltide genereerimiseks on koostatud 25 lähteteksti. Töö tulemusena selgus, et valitud objektide kujutamises oli parim GPT-4o, kuna 31% selle mudeli genereeritud piltidest olid korrektsed. DALL·E 3 genereeritud pildid olid õiged vaid 12% juhtudest ning Stable Diffusioni piltidest

11%. Kuna saavutatud tulemused olid vahemikus 11%–31%, siis saab väita, et GPT-4o, DALL·E 3 ja Stable Diffusion 3.5 ei ole mudeleid maailma kohta integreerinud.

Võtmesõnad:

Tehisintellekt, pildigeneraator, mudelid maailma kohta (ingl *world models*)

CERCS: P176 Tehisintellekt

Contents

Introduction 5

1 Background 7

1.1 World Models 7

1.2 Text-to-image Models and How They Work 9

1.3 Previous Research 10

2 Methods 13

2.1 Text-to-image Models 13

2.1.1 GPT-4o 13

2.1.1 DALL·E 3 14

2.1.2 Stable Diffusion 3.5 14

2.2 Data Collection 15

2.3 Prompt Engineering 16

2.4 Evaluation 17

3 Systematic Analysis 19

3.1 GPT-4o 19

3.2 DALL·E 3 22

3.3 Stable Diffusion 3.5 24

3.4 Comparison 27

4 Discussion 30

Conclusion 32

References 34

Appendices 37

I. Prompts and Points 37

II. Images Generated by GPT-4o, Stable Diffusion 3.5 and DALL·E 3 39

III. License 41

Introduction

In recent years, artificial intelligence (AI) tools such as ChatGPT and various text-to-image generators like DALL·E have become available to the public and have quickly become incredibly popular. New and better versions of these tools are currently being developed - a good example is GPT-4o's image generation, which was released in March 2025 while this thesis was being written [1]. The main purpose of text-to-image models is to generate accurate images that are based on given prompts.

So far, machines have been trained on large image datasets, and a general hope is that larger datasets and more computing power will be the key to better models that will eventually reach the level of human intelligence [2]. But will these machines really improve if they do not apply the core concepts of the human mind? For example, is it possible to create realistic images without understanding natural intelligence and without incorporating world models?

Mental models are a concept of the human mind that define how people perceive and interact with the world around them [3]. In AI and machine learning, the term "world models" is used instead of mental models. The concept of world models stresses the need for AI, including LLMs and text-to-image models, to incorporate the concepts of the human mind. Although the field of AI has recently made huge improvements and may outperform the human mind in many areas, humans still defeat machines in "solving a range of difficult computational problems, including concept learning, scene understanding, language acquisition, language understanding, speech recognition, and so on" [2]. Previous research on AI and its ability to understand relations [4], illusion-illusions [5] and mental model updating [6] has shown that many AI tools have not yet incorporated the concept of world models. B. M. Lake et al. believe that until machines incorporate the concepts of the human mind, humans will continue outperforming AI in the previously referred fields [2].

This research aims to determine whether text-to-image models have integrated world models and to compare their performance in image generation. The study focuses on objects that exhibit a specific logic or pattern. As prompts are one of the key elements of this study, another goal is to identify the best practices in creating them.

Since previous studies have primarily concentrated on relations [4] and topics that, to a certain extent, are more complex, such as illusion-illusions [5] and model-based reasoning [6], this study will focus on simple objects. The text-to-image generators studied are GPT-4o,

DALL·E 3 and Stable Diffusion 3.5 (SD 3.5). For this analysis, 25 prompts were constructed. All selected text-to-image models share a common feature: they generate four images per prompt to give the models more opportunities to succeed. In total, 300 images were generated and analysed between March and April 2025. Ultimately, a conclusion will be drawn on how accurately each text-to-image model creates images that exhibit a logic or pattern, i.e. their ability to demonstrate the use of world models.

This thesis is divided into three main chapters. Chapter 1 provides an overview of world models, text-to-image generators, and previous research. Chapter 2 introduces the methodology, the selected text-to-image models, the data collection process, and the creation of prompts. Chapter 3 presents a systematic analysis of the image creation of GPT-4o, DALL·E 3 and SD 3.5. Appendix I gives an overview of prompts and how many points each text-to-image model received for each prompt. Appendix II presents all 300 images created by text-to-image models

1 Background

This chapter will introduce world models and then provide an overview of text-to-image models and their operational mechanisms. It will also address relevant aspects of previous research, thereby supporting the significance of the present study.

1.1 World Models

Humans are able to naturally acquire knowledge about the world around them, often without significant effort. The notion that both humans and animals engage with mental models was initially introduced by the Scottish psychologist Kenneth Craik in 1943 [7]. The cognitive processes underlying human decision-making are influenced by the mental models that individuals develop through their observations [8]. In essence, mental models determine the manner in which humans comprehend and respond to their environment. Moreover, mental models must demonstrate adaptability to changing circumstances [9] and undergo a process of mental model updating when necessary [10]. Model updating becomes crucial when an error occurs between new data and an existing mental model, prompting either modification of the existing model or the construction of a new one [10].

In AI-related research, the term “world models” is used instead of mental models. Yildirim and Paul [3] have stated: “World models are structure-preserving, behaviorally efficacious representations of the entities and processes in the real world [11], including objects with 3D shapes and physical properties [12], scenes with topological relations and navigable surfaces [13], and agents with beliefs and desires [14]”. Contrary to humans, machine learning models require immense training and are still not capable of matching the cognitive abilities of the human mind [15, p. 2]. For instance, the concept of a die – a six-sided cube with a distinct number of dots on each side, ranging from 1 to 6 – is widely recognised. As illustrated in Figure 1, the text-to-image model DALL·E 3 demonstrates a lack of basic knowledge regarding the die, generating an image of a die with five dots on each side.

Figure 1. A die created by text-to-image model DALL·E 3. Prompt: “Create an image of a die. Make sure that the objects depicted in the image are correct, clear and fully visible”.

The fundamental difference between the human mind and AI lies in the manner in which information is processed [15, p. 3]. Specifically, large language models (LLMs) receive information in written form, whereas humans learn through interacting with the world. LLMs do not have the same level of knowledge as humans – they acquire their understanding through a process of action and training [2]. Although in 2022 LLMs received information from written text only, they now also receive information from audio and images [16].

Moreover, according to M. Bennett [17, pp. 356–357], “it might be inevitable that continuing to scale up these language models by providing them with more data will make them even better at answering commonsense and theory-of-mind questions. And incorporating other modalities like audio and images directly into these models”. M. Bennett argues that despite all the additional modalities, LLMs will not reach human intelligence without world models [17, p. 357].

According to Yann LeCun [15, p. 2], the difference between the abilities of the human mind and those of AI systems can be attributed, in part, to world models. To imitate the cognitive capabilities of the human mind, machines must incorporate world models. Until now, the focus of LLMs has been on immense training with large datasets, but recognising the critical role of world models could shift the future of LLMs. Yann LeCun has stated: “The answer may lie in the ability of humans and many animals to learn world models, internal models of how the world works” [15, p. 2]. Moreover, it is hypothesised that integrating “a set of core cognitive ingredients” by deep learning will result in a positive outcome for the subjects concerned [2].

Although the present chapter has thus far focused mainly on LLMs, the primary objective of this research is to assess the efficacy of text-to-image models in generating visual representations of objects based on world models. Previous research has demonstrated that DALL·E 2 and DALL·E 3 experienced challenges in generating images that accurately represent relations, indicating a lack in world models [4, 18]. B. M. Lake et al. suggest that the use of larger datasets and enhanced computing capabilities will not decrease the shortcomings of AI, and that this approach is not sustainable [2]. Consequently, to generate more accurate depictions, there is a need for machine learning systems, including LLMs and text-to-image models, to incorporate world models.

1.2 Text-to-image Models and How They Work

According to Cloudinary [19], text-to-image models are a subcategory of AI systems that facilitate the generation of images from textual input provided by the user. These tools are gaining popularity due to their efficacy in the domains of digital art, graphic design and visual representation in computer games [19]. To better understand the operational mechanisms of these machines, a simplified and generalised overview of the functionality of text-to-image models is hereby provided.

Text-to-image models employ deep learning algorithms, for example, convolutional neural networks (CNNs), autoencoders (AEs) [20], transformers [21] and diffusion models [22]. CNNs are based on linear algebra principles that facilitate image classification, object recognition and pattern identification. CNNs consist of three layers, with increasing layers representing increased complexity and the ability to identify more extensive portions of the image [23]. AEs have been shown to “encode input data down to its essential features, then reconstruct (decode) the original input from this compressed representation” [24]. The transformer model, such as GPT-4o [16], has exhibited a self-attention mechanism [21]. Due to the self-attention mechanism, the transformer can differentiate the relations between components of an input [21].

DALL·E and Stable Diffusion use diffusion models [22]. According to IBM [22], diffusion models constitute a class of generative models that find primary application in image generation and other computer vision tasks. Diffusion-based neural networks undergo training through a process of deep learning, whereby samples are propagated with random noise in a gradually “diffusing” manner. Subsequently, this diffusion process is reversed to

yield the generation of high-quality images. The diffusion process consists of three main stages: forward diffusion, reverse diffusion, and image generation (see Figure 2) [22].

Figure 2. The diffusion process [25].

It is important to note that neural networks, in isolation, are not capable of accomplishing image generation. According to Cloudinary [19], text-to-image models require extensive training with large datasets consisting of text-image pairs. That helps neural networks to differentiate the relationships between images and textual descriptions. For example, when the model encounters a description such as "a red apple on a wooden table," it learns the visual characteristics of both the apple and the table. When a user inputs a prompt during the generation phase, the model will use the learned database to generate a new image that matches the prompt [19].

1.3 Previous Research

As mentioned earlier, there is a link between the shortcomings of machine learning models and their lack of applying world models. The human mind has no problem understanding the difference between the words on and under, and relationships in general [4]. A human can imagine an image where the sock is under the dog or on the dog. But as studies have shown, this may not be an easy task for AI. Some studies show that text-to-image models seem to not have incorporated world models.

Research done by C. Conwell et al., focusing mainly on DALL·E 2, showed how image generation models fail to understand basic relations, whereas humans understand relations from an early age. Therefore, they believe that DALL·E 2 has not mastered the kinds of representations that help humans perceive the world [4]. Further research by C. Conwell et al. showed that, in addition to basic relations, it is difficult for DALL·E 3, an upgraded version of DALL·E 2, to understand relations when combined with negations or numbers [18].

Research done by T. Ullman has identified an error between vision language models and illusions. According to T. Ullman's explanation, an illusion-illusion is an image that is similar to an illusion, but in essence is not. The experiment demonstrates that vision language models, such as GPT-4o, Calude 3, Gemini Pro Vision, confuse illusion-illusions with illusions. The hypothesis is put forward that if a machine and a human are deceived by the same illusions, it could be indicative of the machine's processing algorithm exhibiting similarities to that of humans [5]. However, if a machine learning model is also deceived by illusion-illusions, it prompts us to reconsider how the original illusion is processed [5]. Therefore, based on these studies, it can be concluded that current vision language models do not have the same perception of the world as humans.

B. Puppart et al. have examined the model-based reasoning of humans and LLMs [6]. They ran a study with a game where the objective was to create a level of difficulty that is straightforward for human players but challenging for LLMs. Despite the game's challenging nature, with a mere 31.6% of humans successfully completing it, humans exhibited a marked superiority over all LLMs. The LLM with the highest performance reached a task rate of 5%. This study supports the hypothesis that humans exhibit superior performance in tasks that demand mental model updating and active engagement in comparison to LLMs [6].

E. Murphy et al. have examined the capacity of text-to-image models to generate images based on sentences used in comprehension tests for children aged 2-7 years. The prompts used in this study are considered to “represent fundamental components of grammatical knowledge”. The study revealed that both text-to-image models, DALL·E 2 and DALL·E 3, could not “match the semantic accuracy of children, even at the youngest age” [26]. Furthermore, as researchers have examined relations, illusion-illusions, mental model updating and visual representations of basic sentences, it may be questioned whether current text-to-image models are capable of creating images of simple objects that are based on world models. As demonstrated by the previous research, a gap exists between AI systems and the human mind.

This research aims to identify whether text-to-image models are able to visually depict objects that follow some specific logic or rules. This study focuses on geometric shapes, musical instruments and objects that adhere to a universal standard, such as an alphabet or computer keyboard. In contrast, without specific rules governing the appearance of objects, text-to-image models operate with greater creativity, making the evaluation of results

challenging. For example, there are no rules for the number of windows a house should have or the number of pages in a book. In comparison, a strict system exists for the arrangement of the dots on the die, as well as the sequence of the letters in the alphabet.

In conclusion, many studies have shown that text-to-image and vision language models have not integrated world models. Previous research has focused on more complex subjects, such as illusion-illusions and relations. Present research focuses on simple objects that follow a logic or specific rules. The results of this study will illuminate whether text-to-image models have something akin to world models or not.

2 Methods

The following section will provide an overview of the three text-to-image models that are the focus of this research. Furthermore, the methodology of data collection and systematic analysis will be outlined.

2.1 Text-to-image Models

The text-to-image models analysed in this research are GPT-4o's image generation, DALL·E 3 and SD 3.5. GPT-4o's image generation has been developed by OpenAI and was launched in 2025 [1]. DALL·E 3 has also been developed by OpenAI and was launched in 2023 [27]. SD 3.5 has been developed by Stability AI and was launched in 2024 [28].

2.1.1 GPT-4o

According to IBM [16]: "GPT-4o is a multimodal and multilingual generative pretrained transformer model released in May 2024 by AI developer OpenAI". It is the newest LLM by OpenAI and is particular for receiving information from multiple sources – text, audio, image and video [16]. As of 2025, GPT-4o has an integrated image generation capability. OpenAI claims that GPT-4o's image generation has, among other things, been demonstrated to excel in precisely following prompts. GPT-4o's image generation has been trained with images and text with the aim to make the relations between images and text understandable as well as interrelationships between images with each other [1]. GPT-4o's image generation is accessible via ChatGPT-4o.

Moreover, human resources have been used in the development of GPT models [17, p. 355]. According to M. Bennett [17, pp. 354–355], OpenAI identified the mistakes of GPT-3 on "commonsense and reasoning questions" and therefore started to train GPT-4 on these questions. Even more, reinforcement learning from human feedback was used. So when GPT-4 answered incorrectly, it was punished and when it answered correctly, it was rewarded. He has stated: "They even pushed the GPT-4 to answer certain questions in specific ways to improve its performance. For example, OpenAI trained GPT-4 to think about commonsense questions by writing out each step, a trick called chain-of-thought prompting. […]. By training GPT-4 to not just predict the answer, but to predict the next step in *reasoning* about the answer, the model begins to exhibit emergent properties of thinking,

without, in fact, *thinking* – at least not in the way that a human thinks by rendering a simulation of a world" [17, p. 355]. Based on this information, it is evident that reinforcement learning from human feedback was also used in the development of GPT-4o and its image generation.

2.1.1 DALL·E 3

DALL·E was launched in 2021 by OpenAI and was the first model capable of creating images based on textual descriptions. DALL·E is a neural network developed to generate images based on the prompts provided by the user [29].

DALL·E 3 was released in 2023 and is a prominent example of a text-to-image model that has gained significant popularity on a global scale. This is the latest version of DALL·E, which was also utilised by ChatGPT-4o until the release of GPT-4o's image generation. Compared to earlier models, the images generated by the DALL·E 3 model are more accurate.. OpenAI has declared that DALL·E 3 is distinguished by its remarkable precision in following the provided description [27]. Since ChatGPT-4o now has its own image generation, DALL·E 3 is accessible via a subfeature within ChatGPT-4o [1].

2.1.2 Stable Diffusion 3.5

SD 3.5 was released in October 2024. SD 3.5 combines different models – SD 3.5 Large, SD 3.5 Large Turbo and SD 3.5 Medium [28]. Each time an image is generated, SD 3.5 evaluates which of the models it will use for a specific prompt. Stable Diffusion's image generation works on a system of credits. Each image that is generated is worth a specific amount of credits because each generation is different and "requires different levels of compute resource so there is variation in the number of credits used" [30]. SD 3.5 is accessible via Dream Studio Beta [31].

SD 3.5 can simultaneously generate multiple images, and the images vary in both style and content. However, a potential drawback of the model is its relative lack of emphasis on aesthetics. Stability AI has declared SD 3.5 to be a market leader because its image quality is comparable to models that use much larger datasets [28].

2.2 Data Collection

This study analyses three text-to-image models: GPT-4o's image generation, DALL·E 3 and SD 3.5. Twenty-five objects have been selected for the purpose of being the primary focus of engineered prompts. The main objective is to ascertain the accuracy of these models in generating images of objects based on world models. The objects in question were either proposed by the supervisor or chosen by the author. The objects were selected on the basis of exhibiting a specific pattern or logic.

Objects belonging to Categories 1-3 are classified according to specific criteria, such as musical instruments, games and sports, and geometric shapes. For instance, Category 2 focuses on music instruments, given that each musical instrument has its own set of rules. Categories 4 and 5 are characterised by the presence of more arbitrary objects, which do not conform to a concrete thematic category. Yet the objects have been selected on the basis of a specific logic or pattern. The complete list of objects is presented in Table 1.

Table 1. Prompt objects.

Category 1	Category 2	Category 3	Category 4	Category 5
Die	Recorder	Triangle	Calendar	Clock
Chessboard	Violin	Pentagon	Ruler	Receipt
Bowling pins	Guitar	Hexagon	Calculation	Thermometer
Football pitch	Harp	Heptagon	Latin alphabet	Speedometer
Basketball court	Piano keyboard	Octagon	US keyboard	Feature phone

Each text-to-image model has been assigned the task of generating images of these objects. Text-to-image models are asked to create 4 images per prompt so they would have more opportunities to succeed and the conclusions made in this study would be more objective. Each model generates a total of 100 images. Consequently, the dataset consists of 300 images in total (see Appendix II). The next section will provide an explanation of prompt engineering.

2.3 Prompt Engineering

In this chapter, the process of prompt engineering is explained, as prompts are of crucial importance. The vast majority of constructed prompts follow a similar logical framework where the first sentence specifies the object to be created and the second sentence tells the text-to-image model that the depicted objects have to be correct and clear. There are a few prompts with minor differences that will be introduced further. It is important that these prompts are not excessively specific, yet they must not be too vague or simplistic.

The following example is provided to illustrate the type of generated prompt that was used in the present study:

1) “Create an image of a receipt. Make sure that the objects depicted in the image are correct, clear and fully visible”.

First, the text-to-image model is asked to create an image of a specific object. The second sentence has been included to inform the text-to-image model of the expectation that the objects should be correct, clear and fully visible. Otherwise, it could be argued that the text-to-image model has not been instructed to be correct. There is a tendency for text-to-image models to depict objects indistinctly or to create images that show only a fraction of the object. Therefore, a decision was made to incorporate these additional details in the prompt.

The following are examples of the more precise prompts used in this study:

1) “Create an image of a clock displaying the time 18:30. Make sure that the objects depicted in the image are correct and clearly and fully visible”.

2) “Create an image of the Latin alphabet. Make sure that the objects depicted in the image are correct and clearly and fully visible”.

In the first prompt, the specification of time is important, because at first it may seem that text-to-image models can depict correct clocks. However, this leads to another question whether they have the capacity to represent a correct time, given its direct correlation to world models. For instance, it appears that the DALL·E 3 has been mainly trained on images of a clock with the clock hands pointing to 2 and 10. If the text-to-image model is asked to create an image of a clock displaying the time 18:30, the results show images where the clock

hands still point to 2 and 10. As clock hands tend to point in a single direction, it was decided by the author that the prompt would also specify the time, thereby giving a more comprehensive understanding about the abilities of text-to-image models.

In the second prompt, the specification of details is important. For instance, it is necessary to state that the model is expected to depict the Latin alphabet, as opposed to a random sequence of characters. In the absence of a specification, the model could generate a Greek alphabet, which would consequently make the evaluation process more difficult. The necessity of specification is also applicable to multiple other objects.

The evaluation of the generated images is the main focus of this study. The subsequent chapter will provide a detailed overview of the systematic analysis and evaluation process.

2.4 Evaluation

A table (see Appendix I) has been constructed to assess the performance of the studied text-to-image models. Given that each model generates 4 images per prompt, the accuracy of the images has been evaluated in a four-point system (see Table 2).

Table 2. Four-point system.

Points	Explanation
0	None of the images are correct.
1	1 of 4 images is correct.
2	2 of 4 images are correct.
3	3 of 4 images are correct.
4	All images are correct.

A four-point system is used in the evaluation process with the aim of reducing subjectivity. An image is correct when it follows the given prompt to detail and there are no shortcomings. For example, an image of a piano keyboard that is partially visible or a guitar with partially blurry strings, is not considered correct.

In addition, each prompt category is analysed in more detail to assess which images are easier and which are more difficult for text-to-image models to create. Moreover, the results of text-to-image models are compared to find similarities and differences in their performance.

Conclusions are made about whether text-to-image models have incorporated world models and which one is most effective.

3 Systematic Analysis

In this chapter, a systematic analysis of the performance of GPT-4o's image generation, DALL·E 3 and SD 3.5 is presented. In total, 300 images will be analysed (see Appendix II). Additionally, the results of the three text-to-image models will be compared.

3.1 GPT-4o

GPT-4o's image generation delivered the best results compared to DALL·E 3 and SD 3.5. Each text-to-image model generated 100 images in total and GPT-4o created 31 correct images (see Appendix II), meaning that 31% of the images were accurate and matched the corresponding prompt.

GPT-4o achieved best results in Category 3 – geometric shapes (see Figure 3). More precisely, 16 out of 20 images were accurate. It excelled at creating images of triangles, pentagons, hexagons and octagons. Interestingly, the only geometric shape it struggled with was a heptagon that should have had 7 peaks and sides but had 5 peaks and sides instead. It is interesting because the only difference between the aforementioned geometric shapes is the number of peaks and sides. If it could easily generate an image of an octagon, why could it not create an image of a heptagon?

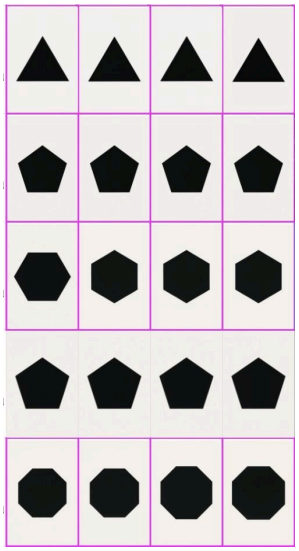


Figure 3. Geometric shapes created by GPT-4o's image generation. Images with borders (magenta) are considered correct.

Second best results were achieved in Category 4 – simple objects that follow a logic or pattern. In this category, 6 out of 20 images were correct. A prompt used in this study was "Create an image of a calculation. Make sure that the objects depicted in the image are correct, clear and fully visible" (see Figure 4). Each of the generated calculations were the same. To have GPT-4o generate an image of a different calculation it should have been specified in the prompt. For example, a prompt such as "Create an image of another calculation. Make sure that the objects depicted in the image are correct, clear and fully visible" was being tested to see whether it would make a difference in the output (see Figure 5).

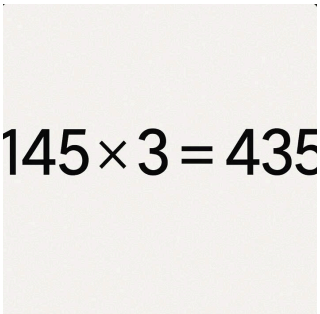


Figure 4. Image of a calculation created by GPT-4o.

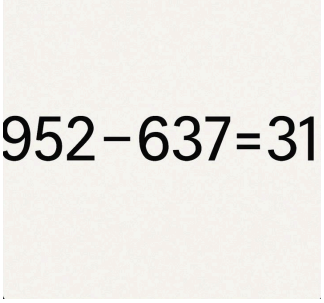


Figure 5. Image of a different calculation created by GPT-4o.

GPT-4o's worst performance was in Category 2 – musical instruments – where it scored 2 out of 20 images (see Figure 6). The main problem with creating images of musical instruments was the instrument's strings that were broken or not attached properly, or were not clear enough to evaluate. Additionally, in some cases, there were small details of the instrument that were missing or inaccurate. In general, the music instruments looked realistic but cannot be considered correct yet. Images where the instrument was only partially visible were not considered accurate.

Figure 6. Musical instruments created by GPT-4o. Images with borders (magenta) are considered correct.

In Categories 1 (games and sports) and 5 (objects that follow a logic or pattern), see Figure 7, the results were not great either. Mostly, there was an error with the overall logic and layout of the objects. For example, the model lacked knowledge of the number of rows and columns of a chessboard or the logic of a thermometer. GPT-4o succeeded in generating 4 accurate images of a football pitch, 1 accurate image of a clock, receipt and feature phone.

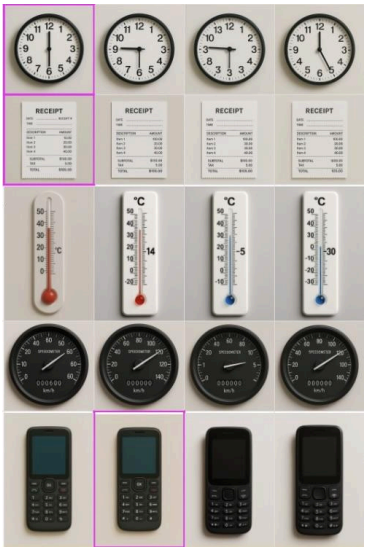


Figure 7. Objects that follow a logic or pattern created by GPT-4o. Images with borders (magenta) are considered correct.

Reaching the score of 31% seems to be the result of training on large amounts of data. Therefore, based on this analysis, it seems that GPT-4o has not integrated world models.

3.2 DALL·E 3

DALL·E 3 created 12 accurate images out of 100, meaning that 12% of the images were correct. Although its performance score is second best, it is far from GPT-4o's 31%.

Like GPT-4o, DALL·E 3 performed best in Category 3 – geometric shapes (see Figure 8). This text-to-image model excelled at creating images of triangles and hexagons but completely failed in visualizing pentagons, heptagons and octagons.

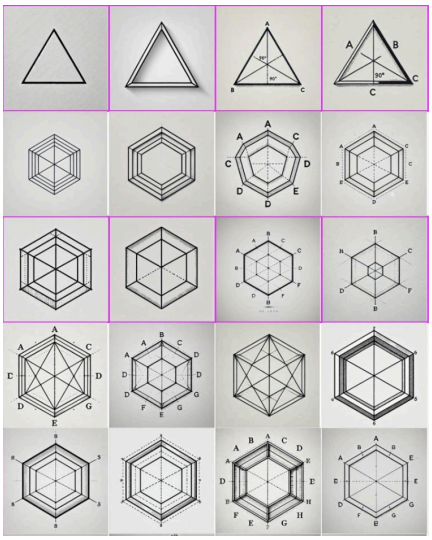


Figure 8. Geometric shapes created by DALL·E 3. Images with borders (magenta) are considered correct.

In Category 1 (games and sports) (see Figure 9), DALL·E 3 managed to accurately depict a football pitch. At the same time these football pitches seem to be rugs. As the football pitch was accurate and it was not specified in the prompt that the text-to-image model was not allowed to combine the prompt object with another object, the generated images were considered correct. The rest of the objects depicted in Category 1 were not accurate. In all studied text-to-image models, there is a general confusion with the number of dots on each side of the die or the number of bowling pins and their layout, or how many baskets does a basketball court have.

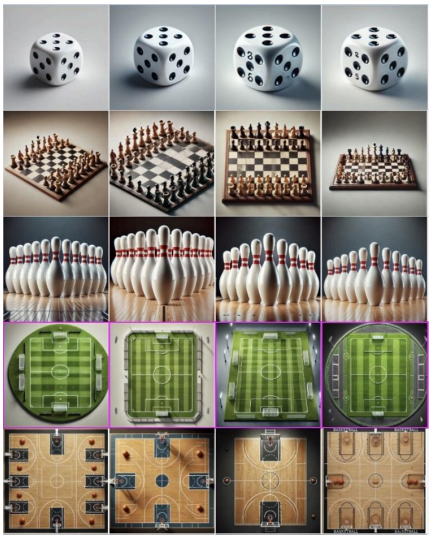


Figure 9. Objects related to games and sports created by DALL·E 3. Images with borders (magenta) are considered correct.

DALL·E 3 completely failed in Categories 2 (musical instruments) (see Figure 10), 4 (simple objects that follow a logic or pattern) and 5 (simple objects that follow a logic or pattern) where none of the images it generated were correct. Mostly, there seemed to be a lack in understanding the general logic of these objects. For example, the problems with music instruments were the same as with GPT-4o – the strings were broken or were not clear enough for evaluation. Other than that, the musical instruments visualized by DALL·E 3 were also quite realistic and similar to GPT-4o's outputs. It is logical as both models were created by OpenAI [1, 27]. In Categories 4 and 5, it is noticeable that there is confusion with numbers and how they should be located on a clock or a feature phone. Additionally, there is confusion with letters and their layout on a keyboard or in an alphabet.

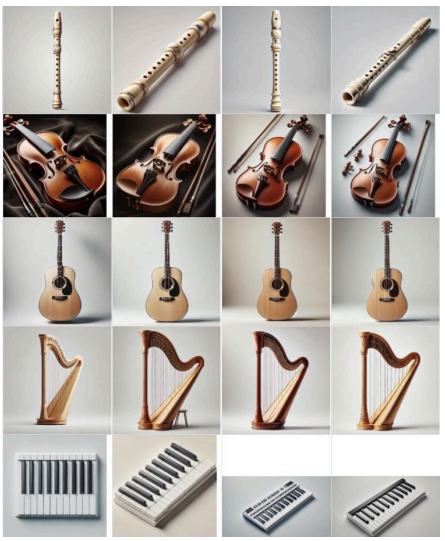


Figure 10. Musical instruments created by DALL·E 3. Images with borders (magenta) are considered correct.

To conclude, 12% is not enough to state that DALL·E 3 has integrated world models. To note, there are no objects that DALL·E 3 can depict and GPT-4o cannot so it is obvious that there has been an advancement with the launch of GPT-4o.

3.3 Stable Diffusion 3.5

SD 3.5's and DALL·E 3's results were almost on the same level being only 1 point apart. SD 3.5 depicted only 11% of the images accurately and was therefore the worst of all three text-to-image models studied. A positive aspect to note is that the style of the images generated by SD 3.5 was very versatile compared to the outputs of GPT-4o and DALL·E 3. Although the style is rather versatile then the aesthetics is lacking, as well as the accuracy of the images.

Again, SD 3.5's best performed best in Category 3 – geometric shapes (see Figure 11). In this category, SD 3.5 scored 11 out of 20 images. Similarly to GPT-4o and DALL·E 3, SD 3.5 was not able to create an image of a heptagon, but was able to visualize an octagon. In addition, generating an image of a pentagon was difficult as it did not create any correct images of it.

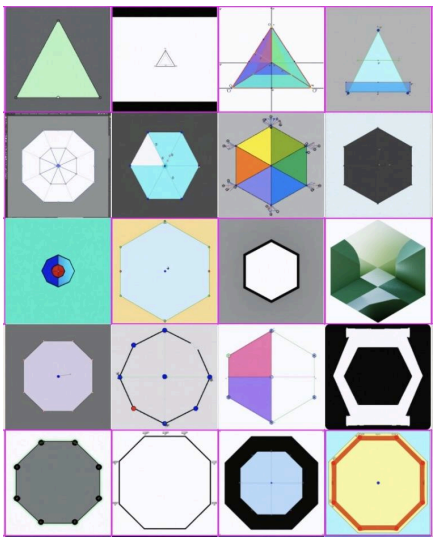


Figure 11. Geometric shapes created by SD 3.5. Images with borders (magenta) are considered correct.

In Categories 1, 2, 4 and 5 SD 3.5 created zero accurate images. The musical instruments were realistic but had the same defects as the musical instruments generated by GPT-4o and DALL·E 3 (see Figure 12).

Figure 12. Musical instruments created by SD 3.5. Images with borders (magenta) are considered correct.

In addition, SD 3.5 also had confusion with numbers and letters, their layout and in many cases the symbols were not readable. They often resembled numbers and letters but were not identifiable. In a few cases, the symbols were missing completely. For example, a calendar created by SD 3.5 had the name of the month but the dates were missing (see Figure 13).

Figure 13. A calendar created by SD 3.5. Prompt: “Create an image of a calendar displaying the month of January. Make sure that the objects depicted in the image are correct, clear and fully visible”.

The benefit of Stable Diffusion is that it can create more than one image at once and it generates the images rather quickly compared to GPT-4o and DALL·E 3. It is all due to the fact that it has smaller datasets. Consequently, the results are worse as well. Some correct (see Figure 14) and failed (see Figure 15) outputs will be presented.

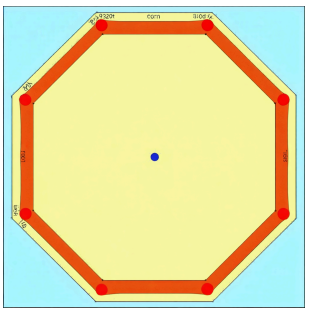


Figure 14. An octagon created by SD 3.5. Prompt: “Create an image of an octagon. Make sure that the objects depicted in the image are correct, clear and fully visible”.

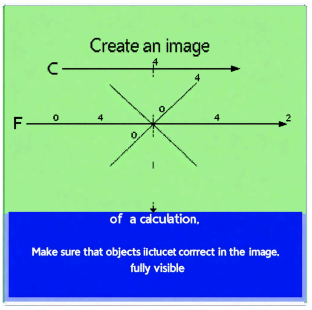


Figure 15. A calculation created by SD 3.5. Prompt: “Create an image of a calculation. Make sure that the objects depicted in the image are correct, clear and fully visible”.

Stability AI says that SD 3.5 is able to understand natural language. Even more, they have stated that: “This ability means that we can describe to the model what we want as we would any other person” [32]. Based on this study, SD 3.5 does not understand natural language and its results in delivering accurate images were poor. In sum, SD 3.5 has not integrated world models.

3.4 Comparison

GPT-4o has proved to be the most effective of the three text-to-image models (see Table 3). Despite its performance being the best it is still far from perfect as 31% is not a good enough score. SD 3.5 demonstrated the poorest performance, achieving a score of 11% and being only 1 point apart from DALL·E 3.

Table 3. Total points for each text-to-image model.

Text-to-image model	Total points
GPT-4o	31/100
DALL·E 3	12/100
Stable Diffusion 3.5	11/100

GPT-4o performed the best in generating correct images. Its performance proved to be substantially better than DALL·E 3 and SD 3.5, with GPT-4o generating 31 accurate images out of 100, as opposed to the 11-12 correct images generated by DALL·E 3 and SD 3.5. However, the score of 31% accurate images is not satisfactory to assert that GPT-4o has really integrated world models. Therefore, it cannot be deduced that any of the text-to-image models have incorporated world models.

DALL·E 3 was the previous model that ChatGPT used for image generation. It is evident that significant progress has been made with the launch of GPT-4o’s image generation – GPT-4o achieved a score of 31% while DALL·E 3 achieved a score of 12%. GPT-4o could depict objects that DALL·E 3 could not, for example, pentagons and octagons. There were no instances where DALL·E 3 created a correct image and GPT-4o did not. The launch of GPT-4o's image generation has definitely had a huge impact on text-to-image generation.

The fact that text-to-image models are able to accurately depict some of the images seems to be still relatively random and not an indication of the text-to-image models’ real capability to depict simple objects that follow a logic. The ability to depict accurate images appears to be

the result of extensive training with large datasets. Interestingly, a text-to-image model is able to accurately depict an octagon, yet encounters difficulties with the representation of a heptagon. In general, all text-to-image models demonstrated an ability to depict the majority of geometric shapes, however, these geometric shapes were always regular. The question of irregular geometric shapes remains unresolved. A preliminary evaluation indicates that GPT-4o exhibits an inability to accurately depict an irregular octagon (see Figure 16), yet demonstrates competence in representing an irregular triangle (see Figure 17). Both of these prompts were tested twice. At first attempt, the triangle was regular and at the second attempt it was irregular. The octagon was regular at both attempts.

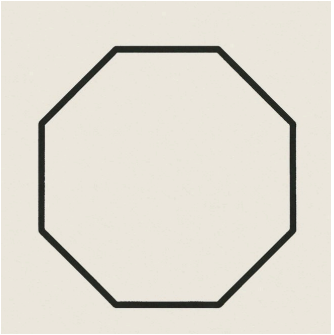


Figure 16. Irregular octagon created by GPT-4o. Prompt: “Create an image of an irregular octagon. Make sure that the objects depicted in the image are correct, clear and fully visible”.

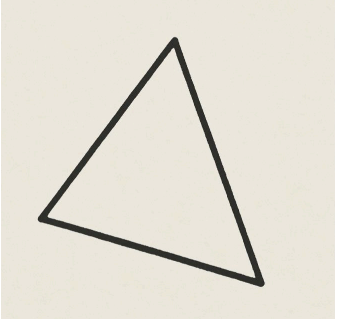


Figure 17. Irregular triangle created by GPT-4o. Prompt: “Create an image of an irregular triangle. Make sure that the objects depicted in the image are correct, clear and fully visible”.

It was also observed that both DALL·E 3 and SD 3.5 exhibited confusion with regard to numbers, letters and their layout. Likewise, GPT-4o encountered difficulties with numbers and letters, but to a lesser extent than DALL·E 3 and SD 3.5. Musical instruments depicted by all text-to-image models were mostly realistic, yet they all exhibited the same defects with some parts being missing or blurry, and broken strings.

The results of this study reveal that text-to-image models GPT-4o's image generation, DALL·E 3 and SD 3.5 have not implemented the concept of world models or have implemented it very poorly. As the results for each text-to-image model were relatively low, it might be an indication that larger datasets and more computing power may not be enough. At the same time, GPT-4o demonstrated a score that was almost three times better than the score of DALL·E 3 and SD 3.5. It might suggest that further improvement in text-to-image generation could be possible using even bigger datasets. However, despite the recent progress it is quite evident that without incorporating world models there will not be an actual breakthrough. Implementing larger datasets and more computing power is possible but not sustainable in the long run to reach human intelligence.

4 Discussion

This thesis shows that text-to-image models have not integrated world models. The core of this analysis were simple objects that follow some logic or rules. Previous research has primarily been conducted on more complex subjects like relations and illusion-illusions. It has shown that text-to-image models and LLMs have not integrated world models. This study, with a focus on simple objects, supports previous findings.

There is a broader discussion about whether the focus should be on using larger datasets and computational capacity in developing text-to-image models or whether there should be a different approach. GPT-4o generated images demonstrate significantly better results than DALL·E 3 and SD 3.5. This indicates that larger datasets and more computational capacity can lead to better results. Yet it is also known that OpenAI developers have used reinforcement learning from human feedback to improve GPT-4o [17, p. 355]. While improving these models with a similar approach to deliver even better results is possible, it seems like a never-ending cycle. There may always be systematic shortcomings that need to be fixed with reinforcement learning from human feedback if the present direction is to be continued.

As this research demonstrates, all text-to-image models delivered poor results even with simple objects. Furthermore, ChatGPT is already struggling with a heavy workload. It often lets its users know that due to a heavy workload users have to wait before they can use its service again. Now that ChatGPT has an integrated image generation the problem with heavy workload could become worse. Moreover, OpenAI recently declared that saying “please” and “thank you” to ChatGPT costs tens of millions of dollars [33]. In this case, larger and larger datasets and more computational capacity seems unrealistic because these machines, and companies that develop these machines, might not be able to manage the workload and expenses. Most probably, there is a need to change direction at some point because there is a lack of resources to manage the heavy workload text-to-image and large language models have been witnessing.

It can be argued that a different approach is necessary for text-to-image models to reach the level of human intelligence. There will possibly be a more powerful text-to-image model than GPT-4o but it still is not enough. Even if larger datasets and more computational power are applied, then these machines will not be independent because they need reinforcement

learning from human feedback. Therefore, it is not possible to reach the level of human intelligence simply by scaling up the current approach.

One factor that has affected this study's results were the prompts and how they were constructed. While general guidelines exist for creating effective prompts, there is no single or correct way to construct them. The prompts used in this study were constructed by the author. The focus was for prompts to be easily readable by humans and to give strict instructions to display the object correctly. But perhaps text-to-image models prefer shorter prompts, which do not need to be readable and comprehensible for humans? It is possible that the results of this research would have been different if the prompts had only consisted of two words. For example, “image chessboard” or “image ruler”.

Future research could examine whether and how shorter prompts would impact the results. How a machine “understands” a prompt also relates to incorporating world models. Therefore, it would be interesting to see whether prompt length affects the results.

Conclusion

The objectives of this study were to: 1) identify whether text-to-image models have integrated world models; 2) create prompts that are neither too precise nor too ambiguous; 3) analyse and compare the output of text-to-image models.

Although AI systems have recently made huge improvements, they still do not beat humans in various fields. The reason might be the differences in how humans and AI gain and process information. Despite all the resources put into improving AI machines, they still lack general skills in many areas. For example, many studies have shown that text-to-image and vision language models have not implemented world models, which is the key to reaching human intelligence.

This study focused on three text-to-image models – GPT-4o's image generation, DALL·E 3 and SD 3.5, all released in recent years. GPT-4o's image generation is the newest as it was launched in March 2025. GPT-4o is a LLM that has an integrated image generation. DALL·E 3 was launched in 2023 and was one of the first significant text-to-image models that gained a lot of popularity. DALL·E 3 was utilised by ChatGPT until the launch of GPT-4o's image generation. SD 3.5 was launched in 2024. It is distinct because it combines different models – SD 3.5 Large, SD 3.5 Large Turbo and SD 3.5 Medium [28].

For this thesis, 25 objects were selected that formed five different categories. The categories were musical instruments, geometric shapes, games and sports, and two categories consisted of simple objects that followed a logic or pattern. The chosen objects were the focus of constructed prompts, and all prompts consisted of two sentences. The first sentence asked the text-to-image model to create an image of a specific object. The second sentence demanded that the objects in the image need to be correct, clear and fully visible.

This analysis shows that the text-to-image models have not implemented world models. The improvements in text-to-image models have mainly been due to using larger datasets and vast amounts of computational capacity. The results for generating correct images were low for all text-to-image models, with the best score being only 31% for GPT-4o. DALL·E 3 created 12% and SD 3.5 generated only 11% of the images correctly. This also means that the way AI machines, including text-to-image models, have been developed and trained so far is not enough to reach the level of human intelligence.

In conclusion, the results of this thesis show that in the studied text-to-image models the concept of world models has not been incorporated. As long as world models have not been incorporated, it is very likely that text-to-image models will not reach the level of human intelligence. The field of AI is rapidly evolving and GPT-4o's image generation has evidently made significant improvements, therefore it is likely there could be surprising developments in the field of text-to-image models in the upcoming years.

References

[1] ‘Introducing 4o Image Generation’. OpenAI. Accessed: Apr. 02, 2025. [Online]. Available: https://openai.com/index/introducing-4o-image-generation/

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, ‘Building Machines That Learn and Think Like People’, Nov. 02, 2016, *arXiv*: arXiv:1604.00289. doi: 10.48550/arXiv.1604.00289.

[3] I. Yildirim and L. A. Paul, ‘From task structures to world models: what do LLMs know?’, *Trends Cogn. Sci.*, vol. 28, no. 5, pp. 404–415, May 2024, doi: 10.1016/j.tics.2024.02.008.

[4] C. Conwell and T. Ullman, ‘Testing Relational Understanding in Text-Guided Image Generation’, Jul. 29, 2022, *arXiv*: arXiv:2208.00005. doi: 10.48550/arXiv.2208.00005.

[5] T. Ullman, ‘The Illusion-Illusion: Vision Language Models See Illusions Where There are None’, Dec. 07, 2024, *arXiv*: arXiv:2412.18613. doi: 10.48550/arXiv.2412.18613.

[6] B. Puppart, P.-H. Paltmann, and J. Aru, ‘Haunted House: A text-based game for comparing the flexibility of mental models in humans and LLMs’, Feb. 12, 2025, *arXiv*: arXiv:2503.16437. doi: 10.48550/arXiv.2503.16437.

[7] K. J. W. Craik, *The nature of explanation*. Cambridge : University Press, 1952. Accessed: Jan. 05, 2025. [Online]. Available: http://archive.org/details/natureofexplanat0000crai/

[8] D. Ha and J. Schmidhuber, ‘World Models’, Mar. 2018, doi: 10.5281/zenodo.1207631.

[9] N. A. Jones, H. Ross, T. Lynam, P. Perez, and A. Leitch, ‘Mental Models: An Interdisciplinary Synthesis of Theory and Methods’, *Ecol. Soc.*, vol. 16, no. 1, 2011, Accessed: May 11, 2025. [Online]. Available: https://www.jstor.org/stable/26268859/

[10] A. Filipowicz, B. Anderson, and J. Danckert, ‘Adapting to change: The role of the right hemisphere in mental model building and updating.’, *Can. J. Exp. Psychol. Rev. Can. Psychol. Expérimentale*, vol. 70, no. 3, pp. 201–218, Sep. 2016, doi: 10.1037/cep0000078.

[11] C. R. Gallistel and A. P. King, *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*, 1st ed. Wiley, 2009. doi: 10.1002/9781444310498.

[12] I. Yildirim, M. Siegel, and J. Tenenbaum, ‘Physical Object Representations for Perception and Cognition’, in *The Cognitive Neurosciences*, 6th ed., D. Poeppel, G. R. Mangun, and M. S. Gazzaniga, Eds., The MIT Press, 2020, pp. 399–410. doi: 10.7551/mitpress/11442.003.0046.

[13] R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, ‘The cognitive map in humans: spatial navigation and beyond’, *Nat. Neurosci.*, vol. 20, no. 11, pp. 1504–1513, Nov. 2017, doi: 10.1038/nn.4656.

[14] J. Jara-Ettinger, H. Gweon, L. E. Schulz, and J. B. Tenenbaum, ‘The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology’, *Trends Cogn. Sci.*, vol. 20, no. 8, pp. 589–604, Aug. 2016, doi: 10.1016/j.tics.2016.05.011.

[15] Y. LeCun, ‘A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27’, 2022, Accessed: Jan. 05, 2025. [Online]. Available: https://openreview.net/pdf?id=BZ5a1r-kVsf/

[16] I. Belcic and C. Stryker, ‘What Is GPT-4o?’. IBM. Accessed: Mar. 11, 2025. [Online]. Available: https://www.ibm.com/think/topics/gpt-4o/

[17] M. S. Bennett, *A Brief History of Intelligence: Evolution, AI, and The Five Breakthroughs That Made Our Brains*. Great Britain: William Collins, 2023.

[18] C. Conwell, R. Tawiah-Quashie, and T. Ullman, ‘Relations, Negations, and Numbers: Looking for Logic in Generative Text-to-Image Models’, Nov. 26, 2024, *arXiv*: arXiv:2411.17066. doi: 10.48550/arXiv.2411.17066.

[19] ‘Text-to-Image Model’. Cloudinary. Accessed: Feb. 17, 2025. [Online]. Available: https://cloudinary.com/glossary/text-to-image-model/

[20] A. Valyaeva, ‘The Comprehensive Guide to Text-to-Image Models’. Accessed: Jan. 05, 2025. [Online]. Available: https://journal.everypixel.com/guide-to-text-to-image-models/

[21] C. Stryker and D. Bergmann, ‘What is a Transformer Model?’. IBM. Accessed: Apr. 11, 2025. [Online]. Available: https://www.ibm.com/think/topics/transformer-model/

[22] D. Bergmann and C. Stryker, ‘What are Diffusion Models?’. IBM. Accessed: Mar. 11, 2025. [Online]. Available: https://www.ibm.com/think/topics/diffusion-models/

[23] ‘What are Convolutional Neural Networks?’. IBM. Accessed: Mar. 11, 2025. [Online]. Available: https://www.ibm.com/think/topics/convolutional-neural-networks/

[24] D. Bergmann and C. Stryker, ‘What Is an Autoencoder?’. IBM. Accessed: Mar. 11, 2025. [Online]. Available: https://www.ibm.com/think/topics/autoencoder/

[25] A. Pan, ‘A Gentle Introduction to Dance Diffusion’, W&B. Accessed: Mar. 12, 2025. [Online]. Available: httpss://wandb.ai/wandb_gen/audio/reports/A-Gentle-Introduction-to-Dance-Diffusion--VmlldzoyNjg1Mzky/

[26] E. Murphy, J. de Villiers, and S. L. Morales, ‘A comparative investigation of compositional syntax and semantics in DALL·E and young children’, *Soc. Sci. Humanit. Open*, vol. 11, p. 101332, Jan. 2025, doi: 10.1016/j.ssaho.2025.101332.

[27] ‘DALL·E 3’. OpenAI. Accessed: Jan. 02, 2025. [Online]. Available: https://openai.com/index/dall-e-3/

[28] ‘Introducing Stable Diffusion 3.5’. Stability AI. Accessed: Jan. 02, 2025. [Online]. Available: https://stability.ai/news/introducing-stable-diffusion-3-5/

[29] ‘DALL·E: Creating images from text’. OpenAI. Accessed: Jan. 02, 2025. [Online]. Available: https://openai.com/index/dall-e/

[30] ‘Plans and Credits’. Dream Studio. Accessed: Apr. 20, 2025. [Online]. Available: https://dreamstudio.stability.ai/plans-and-credits/

[31] ‘DreamStudio’. Accessed: Apr. 20, 2025. [Online]. Available: https://beta.dreamstudio.ai/generate/

[32] ‘Stable Diffusion 3.5 Prompt Guide’. Stability AI. Accessed: May 08, 2025. [Online]. Available: https://stability.ai/learning-hub/stable-diffusion-3-5-prompt-guide/

[33] S. Deb, ‘Saying “Thank You” to ChatGPT Is Costly. But Maybe It’s Worth the Price.’, *The New York Times*, Apr. 24, 2025. Accessed: May 08, 2025. [Online]. Available: https://www.nytimes.com/2025/04/24/technology/chatgpt-alexa-please-thank-you.html/

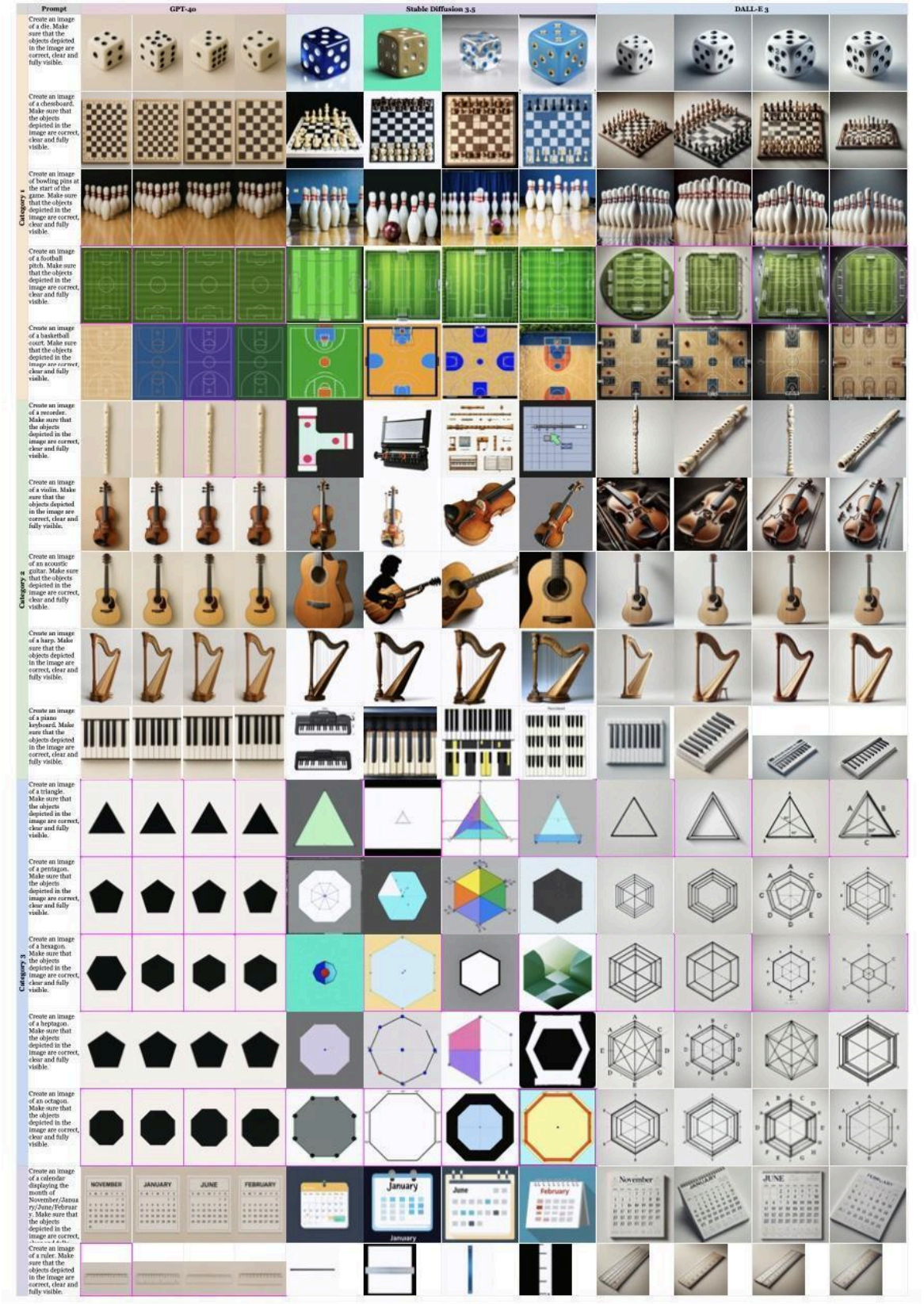
Appendices

I. Prompts and Points

	Prompt	GPT-4o	DALL-E 3	SD 3.5
Category 1	Create an image of a die. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a chessboard. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of bowling pins at the start of the game. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a football pitch. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	4	0
	Create an image of a basketball court. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
Category 2	Create an image of a recorder. Make sure that the objects depicted in the image are correct, clear and fully visible.	2	0	0
	Create an image of a violin. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of an acoustic guitar. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a harp. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a piano keyboard. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
Category 3	Create an image of a triangle. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	4	4
	Create an image of a pentagon. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	0	0
	Create an image of a hexagon. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	4	3
	Create an image of a heptagon. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of an octagon. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	0	4
	Create an image of a calendar displaying the month of November/January/June/February. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0

Category 4	Create an image of a ruler. Make sure that the objects depicted in the image are correct, clear and fully visible.	1	0	0
	Create an image of a calculation. Make sure that the objects depicted in the image are correct, clear and fully visible.	4	0	0
	Create an image of the Latin alphabet. Make sure that the objects depicted in the image are correct, clear and fully visible.	1	0	0
	Create an image of a US keyboard. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
Category 5	Create an image of a clock displaying the time 6:00/9:00/15:30/12:45. Make sure that the objects depicted in the image are correct, clear and fully visible.	1	0	0
	Create an image of a receipt. Make sure that the objects depicted in the image are correct, clear and fully visible.	1	0	0
	Create an image of a thermometer displaying a temperature of +30/+14/-5/-30 degrees Celsius. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a speedometer displaying a speed of 60/120/5/100 km/h. Make sure that the objects depicted in the image are correct, clear and fully visible.	0	0	0
	Create an image of a feature phone. Make sure that the objects depicted in the image are correct, clear and fully visible.	1	0	0
	Total points:	**31**	**12**	**11**

II. Images Generated by GPT-4o, Stable Diffusion 3.5 and DALL·E 3



Note: Images with borders (magenta) are considered correct.

III. License

Non-exclusive licence to reproduce thesis and make thesis public

I, Helena Lindström,

1. Herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, “Systematic Analysis on GPT-4o, DALL·E 3 and Stable Diffusion 3.5. Do Text-to-image Models Incorporate World Models?”, supervised by Jaan Aru.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons’ intellectual property rights or rights arising from the personal data protection legislation.

Helena Lindström
15.05.2025