

# Comparative analysis of the use of instructions for language models and automated metrics for assessing the quality of images generated by GAN models

Yakymiv V. S.<sup>1</sup>, Piskozub Y. Z.<sup>1,2</sup>, Oliarnyk N. R.<sup>1</sup>

<sup>1</sup>*Department of Applied Mathematics,  
Lviv Polytechnic National University,  
12 S. Bandera Str., 79013, Lviv, Ukraine*

<sup>2</sup>*Department of Applied Mathematics,  
Cracow University of Technology,  
24 Warszawska Str., 31-155 Krakow, Poland*

(Received 13 May 2025; Revised 20 September 2025; Accepted 21 September 2025)

This study explores the potential applications of language AI models in combination with Generative Adversarial Networks (GANs) for generating images based on textual descriptions derived from literary works. The effectiveness of various prompt types used to create abstractions was analyzed, and a comparative evaluation of the performance of leading contemporary image generation models – MidJourney, DALL-E, and Stable Diffusion – was conducted. The results indicate that, while language models are capable of producing meaningful abstractions that partially reflect the content of the text, current GAN models do not yet provide the necessary level of semantic correspondence and visual realism. MidJourney demonstrated the highest performance, with DALL-E trailing by only 0.69%, and Stable Diffusion by 11.7%. The evaluation results highlighted the superiority of prompts relying on the generation of generalized abstractions, with the prompt that fully delegated abstraction generation to the language model outperforming the others by 4.52% to 11.18%. In contrast, automated evaluation metrics such as CLIPScore and Inception Score proved inadequate for this specific task. Furthermore, the study discusses the limitations of current GAN training approaches based on “keyword–image” pairs and substantiates the need to enhance generation methods through the use of comprehensive textual descriptions.

**Keywords:** *artificial intelligence; computing; ai-generated images; text-to-image generation.*

**2010 MSC:** 68T50

**DOI:** 10.23939/mmc2025.03.914

## 1. Introduction

A reader may encounter difficulties in fully comprehending the content of a literary work and, consequently, in forming a mental image of the events described. For this reason, illustrations are often incorporated into literary texts. Their primary purpose is to create a specific visual representation that enhances the reader’s perception and emotional engagement with the narrative.

Books that contain images are particularly important for children, especially at an early age. Such reading not only contributes to the development of reading culture and interest in literature among children but also promotes enhanced imaginative abilities [1,2]. Moreover, various studies and surveys indicate that, in the learning process, materials incorporating images are better retained in memory and positively influence the ability to recall key elements of studied educational content [3]. Although this issue has been studied extensively over time, research findings often remain contradictory – for example, some studies suggest that images may serve as a distraction [4]. Nevertheless, other studies demonstrate that the use of illustrations, particularly when their display is delayed until after discussion or reading of the text, can lead to improved comprehension of the meaning conveyed by the written content [5–9].

While contemporary books, particularly those for children, increasingly feature illustrations designed to support the reader's understanding and interpretation of the text, a question arises regarding older literary works, especially classics, for which publishers have not traditionally provided visual accompaniment. This issue is not limited solely to fictional literature but may also be relevant in the context of technical or advertising texts.

This issue can be addressed through Artificial Intelligence (AI) technologies capable of generating images based on predefined parameters. The potential of such technologies is considerable, and the results often exceed expectations. For instance, AI-generated images have won several photography competitions, as the judges believed they were authentic photographs taken by human authors [10,11]. A study was conducted to assess how accurately individuals can distinguish between AI-generated and real images. The findings revealed that participants failed to correctly identify 38.7% of genuine photographs, mistaking them for AI-generated ones [12].

Moreover, in certain cases, this technology can be employed not only as a means of generating specific illustrations but also as a source of inspiration for personal artistic projects [13].

Image generation technology is capable of producing a wide range of visual styles, from pencil sketches to 3D-rendered images, or even illustrations in animation or anime styles, depending on the author's preference. Therefore, such a solution is well-suited for the stated objective.

An application with functionality to convert any text into an image – for instance, a mobile phone application – would be especially useful when reading classical literature. After reading a particular page, the reader could be provided with a visual representation of the events described on that page, thereby enhancing comprehension and engagement with the text.

## 2. Related work

### 2.1. Text-to-image generation

Although AI-based image generation has emerged relatively recently, there is already a large body of research on this topic, with practical applications for solving a variety of problems.

In particular, some studies have shown that AI-generated illustrations can be used in the educational process, as AI can successfully transform abstract or concrete concepts into images [14]. In addition, this technology can influence, inspire, and assist designers in creating product designs [15].

In the field of practical applications of such images, one can single out a comparative analysis study of open AI models for training image recognition based on text prompts to create medical radiological images to improve disease diagnosis and treatment [16]. The result showed that AI models cannot yet reproduce radiological images 100% accurately, but in some cases they can quite accurately reproduce the appearance of some diseases or infections.

Another study examines the role of AI models for biomaterials research in the field of design and construction [17]. This paper, in addition to investigating the creation of biomaterial illustrations, also proposes the use of the ARG (Automatic Referencing Generator) model to add an automatic reference generator to images to guarantee their copyright status.

The following study shows how realistic the generated images can be with a simple text prompt [18]: although such images still have many limitations and understanding of details, the image itself corresponds to the proposed text, and this technology is also developing and improving, offering better and better results.

The aforementioned studies are more focused on analyzing the obtained images or, in general, on assessing realism and compliance with the input parameters, but differ from this one, namely, the goal of this one is to assess the ability of AI to build the main meaning from the pages of books and create an image that corresponds to the description of these pages.

This approach, namely the practical use of AI-generated images, can have negative consequences, both for individual fields and for scientific research. In particular, such illustrations can be presented as real in one's scientific works in order to falsify the results of one's own research [19]. Or, as statistics show, some of the results of AI work are stereotypical in nature and can cause harm if such images are used in a political context [20].

## 2.2. Methods for evaluating generated images

Along with the process of generating images using neural network methods, there is a need for ways to quantify the differences between generated images in a manner that is consistent with human analysis and judgment.

### 2.2.1. Manual image assessment

Images should be evaluated in the same way as any other source, such as an article or book, to determine their quality, reliability, and relevance. The process of evaluating images can involve several steps or types. Visual analysis is the most important way to evaluate an image to understand its meaning. In this case, image characteristics or metadata such as the technical quality of the image, the source of the image, the original context, and textual information about the image play an important role in the evaluation. The following set of questions is suggested for image analysis according to the given characteristic [21]:

Content Analysis:

- What do you see?
- What is the image about?
- Are there people in the image? What are they doing? How are they presented?
- Can the image be viewed from different angles?
- How effective is the image as a visual message?

Visual Analysis:

- How is the image composed? What is in the background and what is in the foreground?
- What are the most important visual elements in the image? How can you determine this?
- How is color used?
- Can the image be viewed in different ways?
- What values are conveyed through the design choices?

Contextual Information:

- What information accompanies the image?
- Does the text change how you see the image? How?
- Is the textual information intended to be factual and informative, or is it intended to influence what and how you see?
- What context does the information provide? Does it answer the questions Where, How, Why, and For Whom was the image made?

Image Source:

- Where did you find the image?
- What information does the source provide about the origin of the image?
- Is the source reliable and trustworthy?
- Was the image found in an image database, or was it used in another context to convey meaning?

Technical Quality:

- Is the image large enough for your purposes?
- Are the colors, lighting, and balance correct?
- Is the image a high-quality digital image, free of pixelation and distortion?
- Is the image in a file format you can use?
- Are there any copyright or other usage restrictions you need to consider?

### 2.2.2. Automatic image assessment

The following metrics are used to automatically evaluate images and assess text relevance and object accuracy:

- Inception Score (IS);
- Frechet Inception Distance (FID);

- Learned Perceptual Image Patch Similarity (LPIPS);
- Structured Similarity Index Metric (SSIM, MSSIM);
- Peak Signal-to-Noise Ratio (PSNR);
- Precision recall;
- CLIPScore;
- Decompositional Alignment Score (DAS);

**Inception Score (IS)** is a mathematical algorithm used to evaluate the quality of images created by generative AI using a generative adversarial network (GAN). This indicator is necessary to automate the process of evaluating generated images, as otherwise, people will need to visually assess each generated image, which is a rather subjective process that may depend on the preferences or biases of the human viewer. This indicator can have a value from zero (the worst result) to infinity (the best result). The IS algorithm is determined by two factors: quality and diversity.

The metric was first considered and proposed in 2016 as an image classification algorithm [22]. Although this metric is quite often mentioned and used for image evaluation, it also has negative aspects. The main disadvantages include the lack of useful guidelines for comparing models, suboptimal metrics, and problems with its application [23].

This indicator also has its limitations:

- small image sizes. The algorithm is only able to work with images of 300x300 pixels;
- limited number of samples. Since AI measures the diversity of images, the limited sample size can lead to an artificially high initial score;
- unusual images. AI can produce an artificially low score if the generated image does not belong to the dataset used to train the algorithm's model.

The algorithm is used to obtain a conditional label distribution  $p(y|x)$ . If the images contain significant objects, they should have a conditional label distribution with low entropy. Also, it is expected that a variety of images will be generated, so the threshold value  $\int p(y|x = G(z))dz$  should have a high entropy. Combining these two requirements, a general mathematical formula for obtaining the IS index will be obtained [22]:

$$\text{IS}(G) = \exp(E_{x \sim p_g} D_{KL}(p(y|x) \| p(y))). \quad (1)$$

**Frechet Inception Distance (FID)** is a metric for quantifying the realism and diversity of images generated by GAN models. The realism score indicates the extent to which generated images, for example of people, look like real, authentic human images. Meanwhile, the diversity metric determines how much the generated images differ from the original to ensure interest and novelty. This metric is well suited for evaluating visual quality and diversity within the same metric. Unlike IS, it can better reflect the similarity of generated images to real ones [24].

Characteristics for which the FID metric can be used:

- GAN score. It shows how well a certain model works in terms of creating realistic images;
- model selection. Helps to compare the performance of different variations or architectures of GAN models;
- hyperparameter configuration. Helps to compare different hyperparameters to configure better performance;
- novelty detection. Detects images that are very different from the original;
- research. Provides a way to compare different GAN models.

The disadvantages of this metric include:

- model limitations due to the limited use of data for its training;
- insensitivity. This metric may not notice small inaccuracies or imperfections in the images;
- preliminary image preparation. To use this metric, all images must be prepared to a single format: all images must be properly scaled, cropped, and normalized;
- subjectivity. The results of the metric may differ significantly from a human assessment;

The FID metric can be calculated using the following formula:

$$\text{FID} = \|\mu - \mu_w\|^2 + \text{tr}(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{\frac{1}{2}}), \quad (2)$$

where  $N(\mu, \Sigma)$  is a multivariate normal distribution estimated by Inception v3 features calculated on real images,  $N(\mu_w, \Sigma_w)$  is a multivariate normal distribution estimated by Inception v3 features calculated on generated (fake) images.

**Learned Perceptual Image Patch Similarity (LPIPS)** is a metric for calculating the structural similarity of high-dimensional images in which pixel values are contextually dependent on each other. This metric is similar to FID in that it uses the internal activations of deep convolutional networks due to their useful ability to represent low-dimensional vectors. Also, LPIPS measures perceptual similarity rather than quality score.

The following steps are used to calculate this metric, namely the calculation of the deep network properties to find the distance  $d_0$  between the reference  $x$  and 2 distorted areas  $x_0$  and  $x_1$  given by the network  $F$  (see Figure 1) [25]:

- calculation of deep embeddings through the transmission  $x$  in  $F$ ;
- normalization of activation in the channel dimension;
- scaling of all channels by the vector  $w$ ;
- calculation of the distance  $l_2$ .

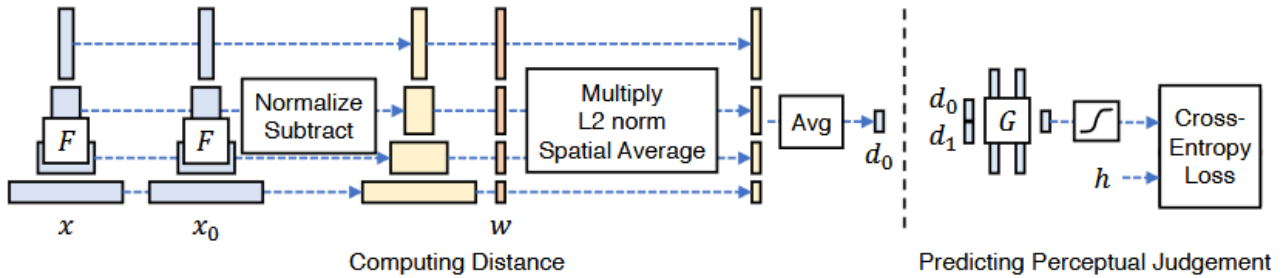


Fig. 1. Schematic diagram of calculating the distance from the network LPIPS.

The formula for calculating the metric is as follows:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \circ (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \right\|_2^2, \quad (3)$$

where  $\hat{y}^l, \hat{y}_0^l \in R^{H_l \times W_l \times C_l}$  is normalization of the feature stack in the channel dimension for layer  $l$ ,  $w^l \in R^{C_l}$  is the vector along which activation is scaled by channels [25].

**Structured Similarity Index Metric (SSIM)** is a new metric developed by the same authors as LPIPS, which mimics the human visual perception system, which has a high ability to identify structural information from a scene. SSIM is designed to determine 3 characteristics of an image: brightness, contrast, structure.

Brightness is determined by the function  $l(x, y)$  and is written as:

$$l(x, y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad (4)$$

where  $C_1$  is a constant to ensure stability when the denominator becomes 0,  $\mu$  is the average value of a given image,  $x$  and  $y$  are the two images being compared.  $C_1$  is given by the formula:

$$C_1 = (K_1 L)^2, \quad (5)$$

where  $L$  is the dynamic range for pixel values (it has value of 255, since the images are standard 8-bit images).

Contrast is defined by the function  $c(x, y)$  and is written as:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad (6)$$

where  $\sigma$  is the standard deviation of a given image,  $x$  and  $y$  are the two images being compared.  $C_2$  is given by the formula:

$$C_2 = (K_2L)^2. \quad (7)$$

The structure is defined by the function  $s(x, y)$  and is written as:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \quad (8)$$

where  $\sigma$  denotes the standard deviation of a given image,  $x$  and  $y$  are the two images being compared, and  $\sigma(xy)$  is defined as:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y). \quad (9)$$

The SSIM metric can be calculated as follows:

$$\text{SSIM}(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma, \quad (10)$$

where  $\alpha > 0$ ,  $\beta > 0$ ,  $\gamma > 0$  denote the relative importance of each metric. This expression can be simplified by assuming that  $\alpha = \beta = \gamma = 1$  and  $C_3 = C_2/2$  [26]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C - 2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (11)$$

However, a separate version of SSIM has also been developed, which, instead of global measurements, measures individual image areas and then calculates the average value for the results [27]. This metric is called the Mean Structural Similarity Index (MSSIM), and as noted, it is more reliable than SSIM and is calculated using the formula:

$$\text{MSSIM}(x, y) = \frac{1}{M} \sum_{j=1}^M \text{SSIM}(x_j, y_j). \quad (12)$$

Another metric for evaluating image quality is **Peak Signal-to-Noise Ratio (PSNR)**. This metric characterizes the peak signal-to-noise ratio in image and video compression tasks, although it is not effective in evaluating images similar to human perception. The metric compares the level of the desired signal with the level of background noise and is expressed in decibels. In cases where the calculation result shows a 1:1 ratio, it means that there is more signal than noise. It is also worth noting that this metric is mainly used in cases where it is necessary to quantify the quality of image reconstruction after compression.

PSNR is calculated using the mean square error (MSE) or  $L_2$  distance and looks like this [28]:

$$\text{PSNR} = 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(\text{MSE}), \quad (13)$$

where  $MAX_I$  is the maximum possible value of image pixels. Also, the higher the PSNR, the higher the quality of image reconstruction.

**Precision recall** is an evaluation metric that can separately and reliably measure quality and coverage in image generation tasks by forming explicit, non-parametric representations of sets of real and generated data [29].

The accuracy metric shows how often an ML model is correct in predicting the target class. In turn, recall shows whether the ML model can find all objects of the target class.

The advantages of using the precision metric include: it works well with unbalanced classes, as it shows that the model is correct in identifying the target class. This metric is useful when the cost of a false positive is high. In such cases, identifying the target class is more important than the possibility of missing some instances. The disadvantages of the metric include the fact that it does not take into

account false negatives, i.e., those cases when the target event is missed. The accuracy is calculated using the following formula:

$$P = \frac{\text{Relevant Items Retrieved}}{\text{Total Retrieved Items}}. \quad (14)$$

The advantages of using the completeness metric include: excellent detection for problems with unbalanced classes, as it focuses on the model's ability to find objects of the target class and for cases where the cost of false positives is high, i.e., if you need to find all objects of the target class, even if this will lead to false positives. Completeness is calculated using the following formula:

$$R = \frac{\text{Relevant Items Retrieved}}{\text{Total Relevant Items in Dataset}}. \quad (15)$$

**Contrastive Language-Image Pretraining (CLIPScore)** is a widely recognized evaluation metric for measuring the similarity between an image generated by a GAN and a corresponding textual description. The main purpose of CLIPScore is to enable models to understand the relationship between visual and textual data and use this understanding to perform tasks such as creating image captions, visual answers to questions. With CLIPScore, you can extract text from images and compare it to a given text in the next step. It has been found to have a high correlation with human judgment.

CLIPScore has the ability to recognize the meaning and context of words in captions and then find the corresponding images in the database. Also, this evaluation model can quickly and accurately match an image caption with a corresponding image, which allows for faster and more accurate image retrieval.

Also, CLIPScore is a cross-modal search model that was trained using 400 million pairs (image and caption) collected from the Internet. The training data represents 500 thousand search queries consisting of common unigrams/bigrams executed in a search engine. For each query, 20 thousand pairs of elements were collected: image and caption. The model has 12 transformation layers and 86M parameters. The text is represented in a similar way: it contains a 12-layer transformer trained on a dictionary of 49 thousand BPE token types.

The metric is defined as follows:

$$\text{CLIPScore}(I, C) = \max(100 * \cos(E_I, E_C), 0), \quad (16)$$

where  $\cos(E_I, E_C)$  is the cosine similarity between the visual CLIP embedding  $E_I$  for the image and the textual CLIPScore embedding  $E_C$  for the signature  $C$ . The score is in the range from 0 to 100, and the closer to 100, the better [30]: images with a score between 0 and 10 can be characterized as rather images generated with a very low level of relation to the input text, from 10 to 30 – with a low, close to average level of relation, more than 30 – contains elements related to the input text and a fairly close text-to-image ratio.

**Compositional Alignment Score (DAS)** is a separate metric that was developed to address the shortcomings of other multimodal models, such as CLIPScore. The disadvantages of such systems are that such models often leave inconsistencies between image and text undetected if the complexity of the input text increases. The DAS metric model, on the other hand, decomposes a complex cue into a set of disjoint statements. In the next step, the VQA model measures the consistency of each statement with the generated images. In the final step, the consistency scores for the different statements are combined to produce a final score.

The difference in the computation compared to CLIPScore is that the computation uses hint-weighting to modify the input embeddings of CLIPScore in the diffusion model:

$$\text{CLIP}(\mathcal{P}) = \mathcal{W}(\mathcal{P}, \{\text{CLIP}(p_i), w_i\}_{i=1}^n), \quad (17)$$

where  $\mathcal{P}$  is the input text,  $\mathcal{W}$  is the hint weighting function,  $p_i$  is the hint corresponding to statement  $a_i$ , and the weights  $w_i$  control the relative weight of different hints  $p_i$  when calculating the overall CLIP embedding for hint  $\mathcal{P}$ .

Also, to increase the strength of the cross-attention of the main noun-subject for each subordinate clause  $a_i$ , the idea of modifying the latent noise  $z_t$  during the backward diffusion process is used,

namely, the modification of the losses for latent updating in the parametric form:

$$z_t = z_t - \alpha \nabla_{z_t} \mathcal{L}(z_t, \{w_i\}_{i=1}^n), \quad \text{where} \quad \mathcal{L}(z_t, \{w_i\}_{i=1}^n) = \sum_i w_i (1 - \max G(\mathcal{A}_i^t)), \quad (18)$$

where  $\alpha$  is the step size,  $\mathcal{A}_i^t$  is the attention map corresponding to the main subject noun in statement  $a_i$ ,  $G$  is the smoothing function, and the weights  $w_i$  determine how much the prominence of different subject nouns in the cue (for each statement) will be increased in the next iteration.

In addition, this approach offers a simpler and more efficient approach to improving text-to-image alignment. Namely, at each iteration  $k \in [1, 5]$  during the refinement process, the overall text-to-image similarity score  $\Omega(\mathcal{I}_k, P)$  and the alignment score at the statement level  $u_i(\mathcal{I}_k, P)$  are calculated. The result of generating image  $\mathcal{I}_{k+1}$  for the next iteration is calculated as follows:

$$\mathcal{I}_{k+1} = \mathcal{D} \left( \mathcal{P}, \{w_i^{k+1}\}_{i=1}^n \right), \quad \text{where} \quad w_i^{k+1} = \begin{cases} w_i^k + \Delta, & \text{if } i = \arg \min_l u_l(\mathcal{I}, P), \\ w_i^k, & \text{otherwise,} \end{cases} \quad (19)$$

where  $\mathcal{D}$  refers to the parameterized diffusion model and  $\Delta$  is the hyperparameter.

This iterative process will be performed until the desired threshold for the overall alignment estimate  $\Omega(\mathcal{I}_k, P)$  is reached. The result of image generation  $\mathcal{I}^*$  at the end of the refinement process is calculated as [31]:

$$\mathcal{I}^* = \arg \max_{\mathcal{I}_k} \Omega(\mathcal{I}_k, P). \quad (20)$$

### 3. Methodology

In this section, a methodology for comparative analysis using several popular AI systems for generating images from text using short descriptions from pages of fiction obtained by AI language models will be presented. Several prepared input prompts for language models will also be used for comparison, to identify and compare them with each other. The study also includes a comparative analysis of the metrics obtained for the automatic evaluation of the generated images of the adversarial networks discussed in the previous section. Since this study aims to investigate the possibility of automatically generating images for any fiction, the availability of original images negates the main purpose of this study. A list of 20 books of classic fiction by different authors was prepared for the study:

- The Count of Monte Cristo by Alexandre Dumas and Auguste Maquet;
- The Sign of the Four by Arthur Conan Doyle;
- Dracula by Bram Stoker;
- A Christmas Carol in Prose; Being a Ghost Story of Christmas by Charles Dickens;
- The Life and Adventures of Robinson Crusoe by Daniel Defoe;
- The Great Gatsby by F. Scott Fitzgerald;
- Metamorphosis by Franz Kafka;
- The Time Machine by H. G. Wells;
- At the mountains of madness by H. P. Lovecraft;
- Moby Dick; Or, The Whale by Herman Melville;
- The Call of the Wild by Jack London;
- Grimms' Fairy Tales by Jacob Grimm and Wilhelm Grimm;
- The Last of the Mohicans by James Fenimore Cooper;
- Pride and Prejudice by Jane Austen;
- Gulliver's Travels by Jonathan Swift;
- A Journey to the Centre of the Earth by Jules Verne;
- The Wonderful Wizard of Oz by L. Frank Baum;
- Alice's Adventures in Wonderland by Lewis Carroll;
- The Picture of Dorian Gray by Oscar Wilde;
- Treasure Island by Robert Louis Stevenson.

Using general random algorithm 5 pages were chosen from each book, receiving at the end 100 test pages.



### 3.1. AI text abstraction extraction

The following open APIs of AI language modeling systems available on the market at the time of writing were used for the study: ChatGPT by OpenAPI with model “gpt-4.5-preview”, Claude3, Gemini.

Other common and popular AI services were not used, as they either do not have an open API for working with the language model or were not available for open use on the market at the time of writing (e.g., Grok or Meta AI models).

To work with the language models, the following abstraction generation samples were prepared to investigate whether a specific command focusing on the length of the abstraction or on the description and key elements of events from the test page would have better results and impact on the quality of image generation:

- P1: You will be provided with a block of text, and your task is to return summarization that describes provided text. Response should not have more than 55 words;
- P2: Choose one scene from the text and prepare summarization of it, describing main characters, scene and what is going on;
- P3: Summarize the following text in a clear and concise manner, preserving all key details and character actions;
- P4: Summarize the following text in no more than 55 words, ensuring that all key events, characters, and motivations are clearly conveyed;
- P5: Provide such summary for some scene from provided text, in the way that it can be used for AI image generation;

### 3.2. Image generation

The next stage of the study is to generate images using competitive generative AI networks using the generated short descriptions. Since this study is a continuation of the previous ones [32, 33], it was decided to use the same models: Dall-e, model “dall-e 3”; Stable Diffusion, version 6; Midjourney, version 6.1.

### 3.3. Image and abstraction scoring

The following scale was proposed to evaluate the received text description (see Table 1).

**Table 1.** Scale for assessing the relevance of keywords or descriptions to the text.

| Text mark           | Number mark | Description  |
|---------------------|-------------|--|
| Not relevant        | 1–2         | The provided keywords or summary have minimal to no connection with the content of the text. There’s little or no overlap in terms of theme or information.            |
| Slightly Relevant   | 3–4         | There are some minor connections between the keywords or summary and the text, but the correspondence is weak. The keywords may touch on peripheral aspects.           |
| Somewhat Relevant   | 5–6         | There is a moderate level of relevance. The keywords or summary capture some aspects of the text, but there are notable gaps or differences.                           |
| Moderately Relevant | 7–8         | The keywords or summary align well with the text, capturing the main ideas and themes. However, there may be some nuances or details that are not perfectly reflected. |
| Highly Relevant     | 9           | The keywords or summary closely match the content of the text. They effectively encapsulate the main points and themes, with only minor variations.                    |
| Perfectly Relevant  | 10          | The keywords or summary perfectly describe the text. Every important detail, theme, and nuance is accurately reflected.  |

The same rating scale was used to evaluate the relevance of the images to the text.

In addition to evaluating the generated images using the obtained abstractions, the raw text from selected pages of fiction was also used to compare whether this approach would have better results and whether the abstraction approach is justified at all.

In evaluating the text and the corresponding image, consideration was given to the extent to which the text contains detailed descriptions of objects, the number of characters, their appearance, and the depiction of events. The assessment also accounted for how much information the text provides to enable a human reader to comprehend the scene's imagery and potentially generate a poster based on the narrative.

### 3.4. Metrics for automation scoring

The following metrics were selected for this study: CLIPScore, Inception Score.

These metrics can provide better automatic evaluation, and are best suited for this study, as some metrics (e.g., FID) require a set of original images to compare with the generated ones.

The range of values for the CLIPScore metric is from 0 to 100. The Torch and Clip libraries with the ViT-B/32 model were used to calculate the metric. The calculation functions return the raw value of the metric obtained from the cosine similarity function. In most cases, to normalize the calculations, the result is multiplied by 100 to obtain a value in the range from 0 to 100. However, since the scale for manual evaluation of images ranges from 0 to 10, multiplication by 10 instead of 100 was used to normalize the CLIPScore metric to bring the results to a single format for comparison with the results of manual evaluation.

The IS metric does not have a fixed maximum score. It is theoretically unlimited in its maximum value, but in real conditions the values are usually in the range from 1 to 10–12, in cases where the generated images have very high values, which are typical for real images or very well generated images. This metric is based on the KL-divergence between  $p(y|x)$ , the probability of the predicted class for each image, and  $p(y)$ , the average probability of classes for all images. The more confidently the model classifies each image ( $p(y|x)$ ), and the more evenly the classes are distributed in the set ( $p(y)$ ), the higher the Inception Score will be. For example, if each image is classified as a unique class with 100% confidence, the IS will be very high. But this does not always mean realism.

Therefore, all the results should be normalized and brought to a single evaluation scale to allow for comparison of results with each other.

## 4. Results and discussions

The main goal of this study is to compare the differences in evaluation results for different samples to identify the dependence of the correctness of the input command for further process improvement and obtaining realistic and correct generated images that would reflect the essence of the scanned fiction text. In addition, the goal is to calculate automated metrics for evaluating the generated images and compare them with human ratings to determine whether such metrics can be used to automate the generation process and consistently evaluate the quality with a possible re-request for generation in case of inadequate quality images.

The evaluation was carried out only by the author of the study, so a certain subjective error in the obtained estimates should be taken into account.

### 4.1. Text abstraction score results analysis

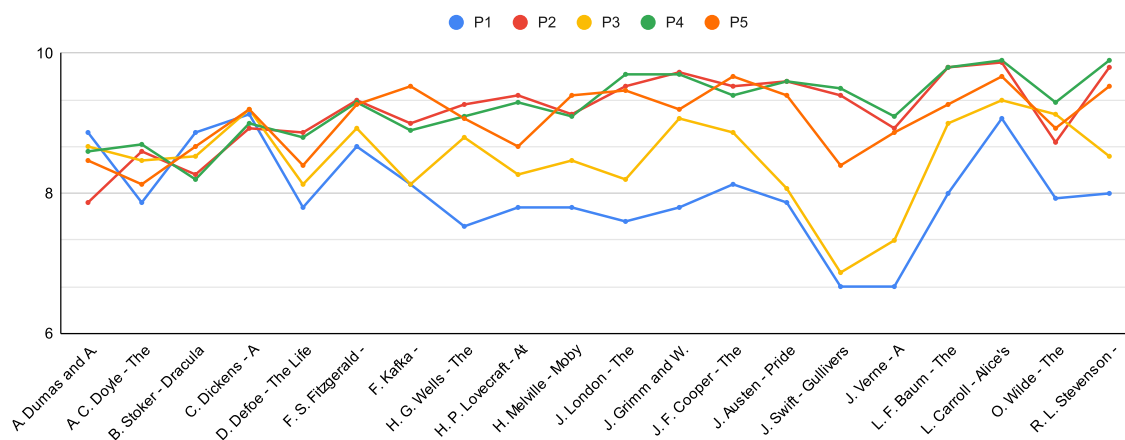
From the data obtained, for each individual book, the average value for each used prompt was calculated depending on the language model, as well as for the images generated by the AI. The average values of abstraction scores grouped by prompt number depending on the language model are presented table (see Table 2).

Based on the results presented in Table 2 and the diagram (see Figure 2), we can see a significant difference between the scores for the given prompts: for most of the books tested, prompts P1 and P3 have the lowest scores. Prompt P1 is a very similar prompt to those used in previous studies [32,33], with the difference that in this case a 55-word limit was added. This 55-word limit was chosen

because some metric algorithms have a limit on the number of words that can be used to check the correspondence of the generated image to the text. Prompt P4 also has a word limit of 55 words, but the evaluation results are much higher than in P1. Prompt P3 instructs the language model to return the main details and actions of the characters in abstraction, just like in P2 and P4. These prompts do not differ too much in the way they give commands to the language model in terms of human perception, but the results show that these differences are important for the language model and have a strong impact on the result. In turn, Prompt P5, which had no restrictions and did not specify the details that should be present in the description, but rather gave the language model the command to return an abstraction that could provide the GAN model with a description to better generate the corresponding image, has quite high evaluation results, on par with P2 and P4.

**Table 2.** Average prompt score by ML.

|                  | Claude3 |     |     |     |     | Gemini |     |     |     |     | GPT-4 |     |     |     |     |
|------------------|---------|-----|-----|-----|-----|--------|-----|-----|-----|-----|-------|-----|-----|-----|-----|
|                  | P1      | P2  | P3  | P4  | P5  | P1     | P2  | P3  | P4  | P5  | P1    | P2  | P3  | P4  | P5  |
| A. Dumas         | 9       | 7.6 | 8.2 | 8.8 | 8.4 | 8.4    | 7.8 | 8.8 | 8.6 | 7.4 | 9.2   | 8.2 | 9   | 8.4 | 9.6 |
| A. C. Doyle      | 7.6     | 8.8 | 8.4 | 8.4 | 8.2 | 7.6    | 8.6 | 8.2 | 8.2 | 7.8 | 8.4   | 8.4 | 8.8 | 8.6 | 8.4 |
| B. Stoker        | 8.8     | 8.4 | 8.4 | 9.2 | 8.8 | 9      | 8   | 8.6 | 8.8 | 8.8 | 8.8   | 8.4 | 8.6 | 9   | 8.4 |
| C. Dickens       | 9.2     | 9   | 9   | 9.4 | 9.8 | 9.2    | 9   | 9.6 | 9   | 8.8 | 9     | 8.8 | 9   | 9.2 | 9   |
| D. Defoe         | 7.8     | 8.6 | 8.8 | 7.6 | 8.8 | 7.6    | 9   | 7.8 | 7.4 | 8   | 8     | 9   | 7.8 | 7.6 | 8.4 |
| F. S. Fitzgerald | 8.4     | 9.2 | 8.8 | 9   | 9.4 | 8.8    | 9.4 | 8.8 | 8.4 | 8.8 | 8.8   | 9.4 | 9.2 | 8.8 | 9.6 |
| F. Kafka         | 8.6     | 9   | 8.2 | 8.6 | 10  | 7.8    | 8.8 | 8   | 8.6 | 8.6 | 8     | 9.2 | 8.2 | 8.8 | 10  |
| H. G. Wells      | 7.4     | 9   | 8.4 | 8.2 | 9.4 | 7.4    | 9.2 | 9.2 | 8.4 | 8.6 | 7.8   | 9.6 | 8.8 | 8.6 | 9.2 |
| H. P. Lovecraft  | 8       | 9.4 | 8   | 7.6 | 9   | 7.6    | 9.2 | 8.2 | 7.8 | 8   | 7.8   | 9.6 | 8.6 | 8.2 | 9   |
| H. Melville      | 7.6     | 9.2 | 8.6 | 8   | 9.8 | 8      | 9   | 8.4 | 8.2 | 8.4 | 7.8   | 9.2 | 8.4 | 8.2 | 10  |
| J. London        | 7.8     | 9.8 | 8.6 | 8   | 10  | 7.4    | 9.6 | 8   | 7.6 | 8.8 | 7.6   | 9.2 | 8   | 7.6 | 9.6 |
| J. Grimm         | 7.6     | 9.8 | 9.2 | 8.4 | 9.4 | 8      | 9.6 | 8.8 | 8   | 8.6 | 7.8   | 9.8 | 9.2 | 8.4 | 9.6 |
| J. F. Cooper     | 7.8     | 9.4 | 9.2 | 7.6 | 9.6 | 8.4    | 9.4 | 8.4 | 7.2 | 9.4 | 8.2   | 9.8 | 9   | 7.8 | 10  |
| J. Austen        | 7.8     | 9.6 | 8.6 | 7.4 | 9.8 | 8      | 9.6 | 7.6 | 7.2 | 9   | 7.8   | 9.6 | 8   | 7.2 | 9.4 |
| J. Swift         | 6.8     | 9.8 | 7.4 | 6.6 | 8.8 | 6.8    | 9.2 | 6.6 | 6.6 | 8.4 | 6.4   | 9.2 | 6.6 | 6.6 | 8   |
| J. Verne         | 6.4     | 9   | 7.4 | 6.6 | 9.2 | 7      | 9.2 | 7.2 | 7.8 | 7.8 | 6.6   | 8.6 | 7.4 | 8   | 9.6 |
| L. F. Baum       | 8       | 9.6 | 9   | 8.2 | 9.6 | 8      | 10  | 9.2 | 8.4 | 8.8 | 8     | 9.8 | 8.8 | 8.8 | 9.4 |
| L. Carroll       | 9.2     | 9.8 | 9   | 9   | 10  | 9      | 10  | 9.8 | 8.8 | 9.2 | 9     | 9.8 | 9.2 | 9   | 9.8 |
| O. Wilde         | 7.8     | 9.2 | 9   | 8.8 | 8.8 | 7.8    | 9.4 | 9.2 | 8.6 | 8.2 | 8.2   | 7.6 | 9.2 | 8.6 | 9.8 |
| R. L. Stevenson  | 7.6     | 10  | 8.2 | 8.2 | 9.6 | 8      | 9.8 | 8.8 | 9   | 9   | 8.4   | 9.6 | 8.6 | 8.6 | 10  |
| <b>AVG</b>       | 8.64    |     |     |     |     | 8.46   |     |     |     |     | 8.68  |     |     |     |     |



**Fig. 2.** Diagram for average prompt score by book.

#### 4.2. Image generation errors analysis

The generation of images by GAN models is accompanied by a problem that can significantly affect the result, or even make it impossible to obtain the result at all. The matter is that AI services have a list of unacceptable words that are used to check queries sent by users. Examples of such unacceptable words include racist, politically incorrect, or swear words.

The API is designed to repeat the request in case of an error to minimize the risk of possible Internet outages, internal system errors, or false alarms (e.g., in the case of DALL-E input word filters).

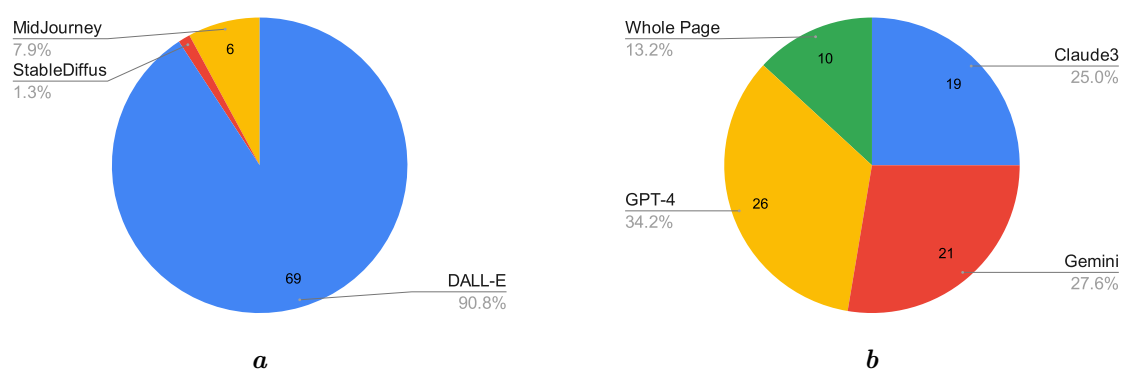
The statistics of the errors obtained during the study are presented in table (see Table 3).

**Table 3.** Number of errors for tested AI services.

|                 | DALL-E | StableDiffusion | MidJourney | SUM |
|-----------------|--------|-----------------|------------|-----|
| Claude3         | 18     | 1               | 0          | 19  |
| Gemini          | 20     | 0               | 1          | 21  |
| GPT4            | 25     | 0               | 1          | 26  |
| Whole Page Text | 6      | 0               | 4          | 10  |
| SUM             | 69     | 1               | 6          |     |

Most errors were obtained when using the GAN model of DALL-E (see Figure 3a). DALL-E also refused to process queries that contained descriptions of bloody scenes. For example, most of these cases occurred when generating images for randomly selected pages from Lovecraft's "In the Mountains of Madness". These results indicate that the DALL-E GAN model is quite risky to use for the task at hand, as the model has limited potential due to the large number of input filters and cannot correctly process queries even from classic fiction.

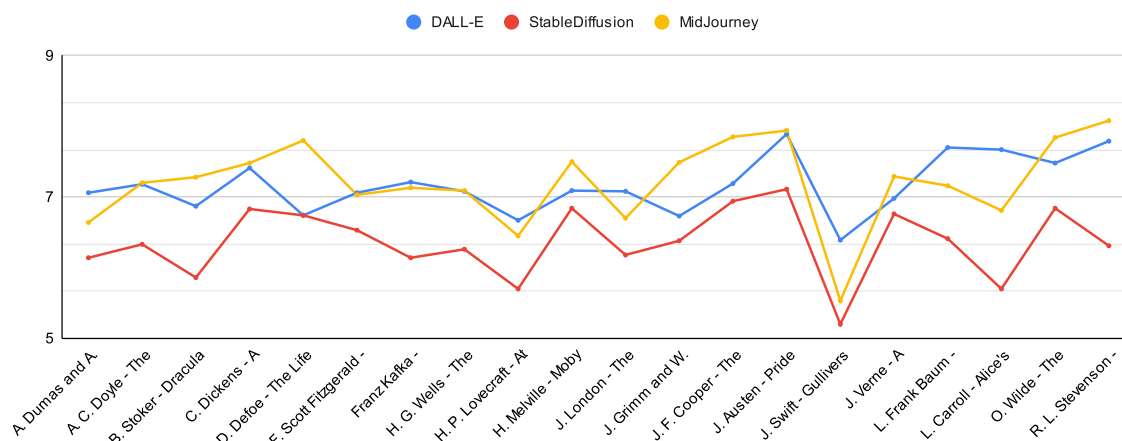
Among the tested language models that generated abstractions, the GPT-4 ("gpt-4.5-preview") model has the largest number of errors in generating images (see Figure 3b), while other models have a small difference in the number of errors. The lowest number of errors was obtained when using the entire text of a fiction page. This result may indicate that with a larger amount of text, the GAN model can select only some elements from the text and ignore incorrect or invalid phrases and words that would otherwise cause the model's API to reject the query. This fact gives the whole-page approach significant advantages over language models.



**Fig. 3.** Number of errors: (a) by GAN model and (b) by language model and using whole page.

#### 4.3. Image generation score results analysis

On the graph showing the dependence of the GAN model scores on the book (see Figure 4)), MidJourney showed the highest scores for almost all the books tested, while Dall-e has similar but slightly worse results, and StableDiffusion has the worst results, with the difference between StableDiffusion and MidJourney reaching two points. The average results of MidJourney are quite high, but still far from the maximum scores, which may indicate that in general GAN models are still in their infancy and have shortcomings that should be improved in future versions of the models.



**Fig. 4.** Average image scores between DALL-E, MidJourney and StableDiffusion to book.

Among other things, it is worth noting that GAN models are trained on word pairs to obtain images. And although the results of using keywords were worse compared to abstractions [32], it can still be concluded that using a textual whole description is difficult for GAN models, because it selects only some words or word pairs for generation, ignoring the entire provided description and content and losing details from the textual description. In particular, the more descriptions of objects or characters the text contains, the more often these objects/characters are combined, swapped or lost.

For example, this abstraction was obtained for the P5 prompt using the Claude3 language model from “At the mountains of madness” by H. P. Lovecraft:

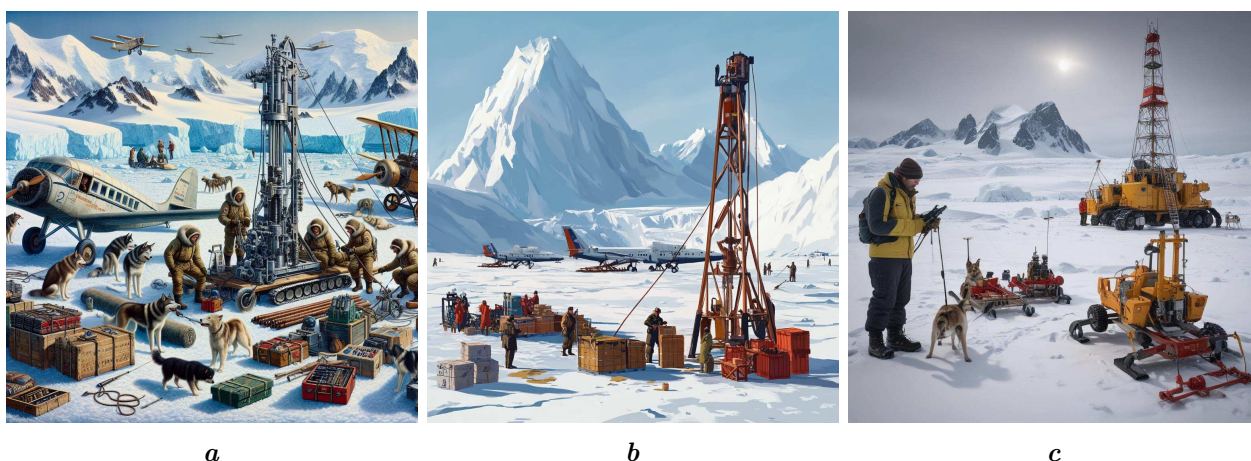
A group of scientists and explorers prepare for an Antarctic expedition. The scene features a mix of cutting-edge technology and traditional exploration equipment:

1. A unique, lightweight drilling apparatus with a steel head, jointed rods, and a gasoline motor.
2. A collapsible wooden derrick stands nearby.
3. Three large dog sledges loaded with equipment, including dynamite, cording, and sectional piping.
4. Four large Dornier aeroplanes, specially designed for high-altitude flying, are parked on the ice.
5. Teams of sled dogs are visible, ready for the journey.
6. In the background, vast ice fields and mountain ranges of Antarctica stretch out.
7. Scientists in cold-weather gear examine maps and equipment.
8. Crates of supplies and scientific instruments are scattered around the camp.

The overall atmosphere is one of excitement and anticipation, with a blend of industrial and natural elements set against the stark, icy landscape of Antarctica.

For the abstraction mentioned above, next images were obtained (see Figure 5), including: StableDiffusion contains only a small part of the described equipment; MidJourney contains additional equipment and 2 aircraft, although the abstraction indicates exactly 4; DALL-E contains more of the specified equipment, but in general the scene is not realistic, overloaded with details, and the number of aircraft and people is incorrect.

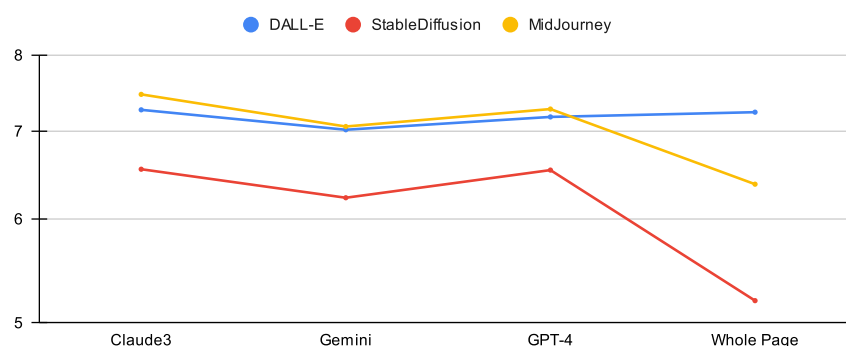
For the summary data of the GAN model versus language model (see Figure 6), we can see an almost uniform relationship between the models. DALL-E and MidJourney show almost the same results for all language models used, except for the use of a whole page of text, where DALL-E showed better results. StableDiffusion, on the other hand, scored uniformly lower. The summarized results of evaluating the dependence of the GAN model on the books under test are presented in table (see Table 4).



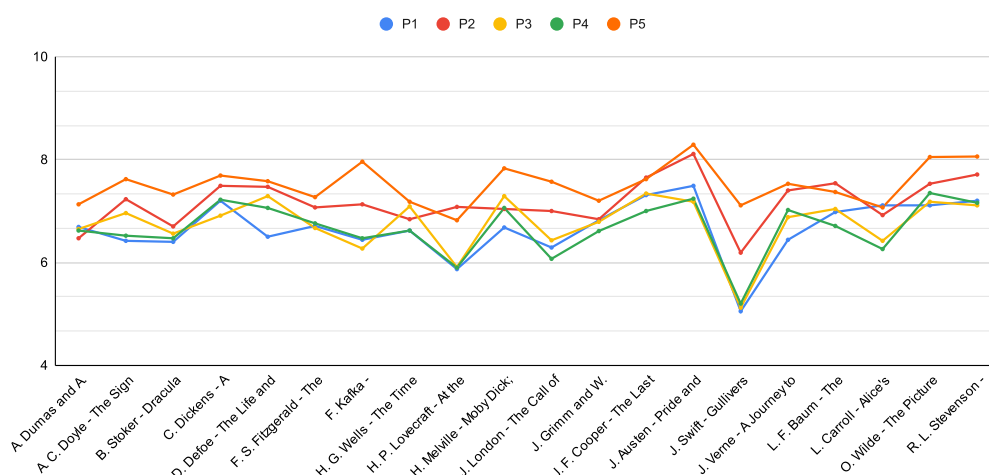
**Fig. 5.** Examples of generated images (a) DALL-E and (b) MidJourney and (c) StableDiffusion from “At the mountains of madness” by H. P. Lovecraft using P5 and Claude3 language model.

The diagrams (see Figures 7–8) contain the average scores for the generated images depending on the prompts for each book tested. These graphs show the highest results for the prompts P2 and P5, which specify the command to extract key characters and events from a single scene of the text (if the selected

page describes several scenes, the model should independently select and describe only one of them) and the command in which the language model should independently prepare a description that could be used as a description for generating images by GAN models. Moreover, it is the P5 prompt that dominates most of the graph, which demonstrates how the language model can independently evaluate which description will be better for the GAN model and choose the right description, better than a human.



**Fig. 6.** Average image scores between DALL-E, MidJourney and StableDiffusion to AI ML.

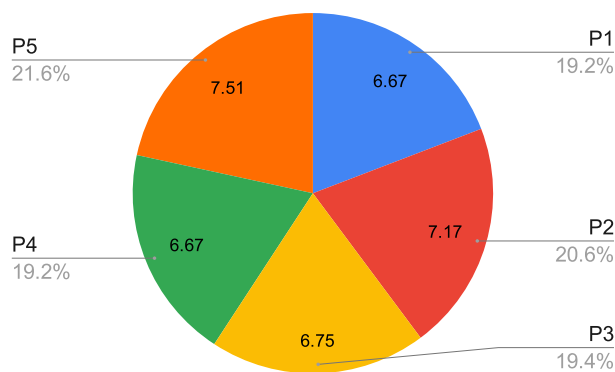


**Fig. 7.** Average image scores between used prompts to book.

**Table 4.** Average score between DALL-E, MidJourney and StableDiffusion to book.

|                       | DALL-E | StableDiffusion | MidJourney |
|-----------------------|--------|-----------------|------------|
| A. Dumas              | 7.06   | 6.14            | 6.64       |
| A. C. Doyle           | 7.18   | 6.33            | 7.2        |
| B. Stoker             | 6.87   | 5.86            | 7.28       |
| C. Dickens            | 7.41   | 6.83            | 7.48       |
| D. Defoe              | 6.74   | 6.74            | 7.8        |
| F. S. Fitzgerald      | 7.06   | 6.53            | 7.03       |
| F. Kafka              | 7.21   | 6.14            | 7.13       |
| H. G. Wells           | 7.08   | 6.26            | 7.09       |
| H. P. Lovecraft       | 6.67   | 5.7             | 6.45       |
| H. Melville           | 7.09   | 6.84            | 7.5        |
| J. London             | 7.08   | 6.18            | 6.7        |
| J. Grimm and W. Grimm | 6.73   | 6.38            | 7.49       |
| J. F. Cooper          | 7.19   | 6.94            | 7.85       |
| J. Austen             | 7.89   | 7.11            | 7.94       |
| J. Swift              | 6.39   | 5.2             | 5.53       |
| J. Verne              | 6.98   | 6.76            | 7.29       |
| L. F. Baum            | 7.7    | 6.41            | 7.16       |
| L. Carroll            | 7.67   | 5.7             | 6.81       |
| O. Wilde              | 7.48   | 6.84            | 7.84       |
| R. L. Stevenson       | 7.79   | 6.31            | 8.08       |
| AVG                   | 7.16   | 6.36            | 7.21       |

The lowest scores on the graph are demonstrated by the prompts P1 and P4, which have a description length limit of 55 words. Therefore, we can conclude that either language models trying to generate an abstraction in a limited number of words result in an incorrect or inaccurate description, or GAN models should use longer descriptions for the purpose.

**Fig. 8.** Aggregated image ratings across used prompts.

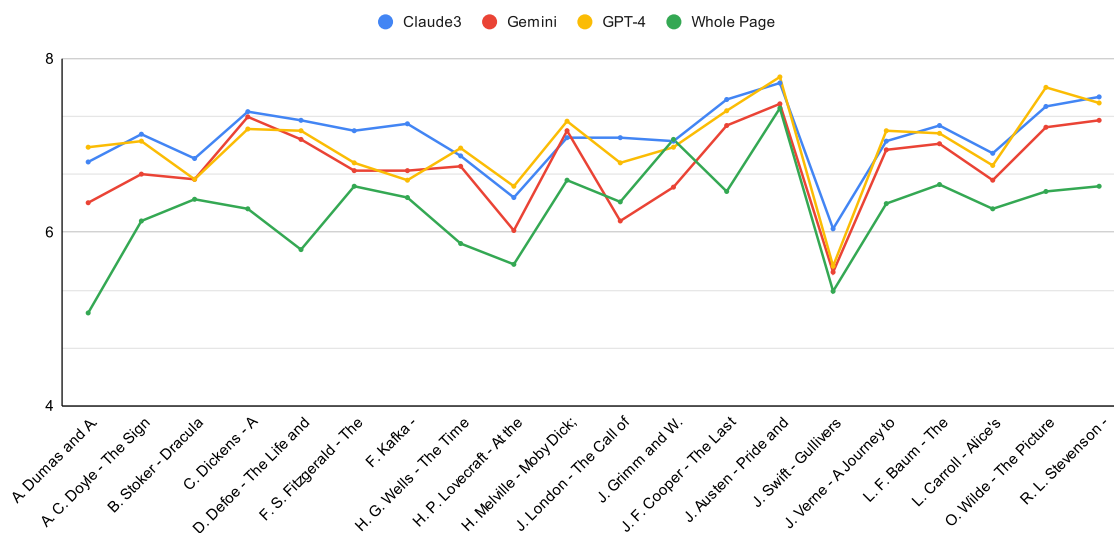
The summary data for image evaluations generated by GAN models grouped by prompts are shown in table (see Table 5). The diagrams (see Figures 9–10) show the average scores for the generated images depending on the language model used for each book tested. These graphs show the highest results for the Claude3 and GPT-4 models, with Claude3 slightly prevailing. The lowest scores were obtained for images that used the entire text page of a work. In general, the graph (see Figure 10) shows fairly comparable values between the models for all the books tested. Therefore, since the difference between the scores of the AI language models used is insignificant, all 3 can be used for further use and research, but Claude3 and GPT-4 are preferred.

The summarized averages for image evaluations generated by GAN models grouped by AI language models are shown in table (see Table 6).



**Table 5.** Average score between used prompts to book.

|                       | P1   | P2   | P3   | P4   | P5   |
|-----------------------|------|------|------|------|------|
| A. Dumas              | 6.69 | 6.47 | 6.65 | 6.62 | 7.13 |
| A. C. Doyle           | 6.42 | 7.23 | 6.96 | 6.52 | 7.62 |
| B. Stoker             | 6.4  | 6.7  | 6.56 | 6.47 | 7.32 |
| C. Dickens            | 7.2  | 7.49 | 6.91 | 7.22 | 7.69 |
| D. Defoe              | 6.5  | 7.47 | 7.29 | 7.06 | 7.58 |
| F. S. Fitzgerald      | 6.71 | 7.07 | 6.67 | 6.76 | 7.27 |
| F. Kafka              | 6.44 | 7.13 | 6.27 | 6.47 | 7.96 |
| H. G. Wells           | 6.62 | 6.84 | 7.09 | 6.62 | 7.18 |
| H. P. Lovecraft       | 5.87 | 7.08 | 5.91 | 5.9  | 6.82 |
| H. Melville           | 6.68 | 6.04 | 7.29 | 7.06 | 7.83 |
| J. London             | 6.29 | 7    | 6.43 | 6.07 | 7.57 |
| J. Grimm and W. Grimm | 6.82 | 6.84 | 6.79 | 6.61 | 7.2  |
| J. F. Cooper          | 7.31 | 7.65 | 7.34 | 7    | 7.62 |
| J. Austen             | 7.49 | 8.11 | 7.18 | 7.24 | 8.29 |
| J. Swift              | 5.05 | 6.19 | 5.12 | 5.2  | 7.11 |
| J. Verne              | 6.44 | 7.4  | 6.88 | 7.02 | 7.53 |
| L. F. Baum            | 6.98 | 7.54 | 7.04 | 6.71 | 7.37 |
| L. Carroll            | 7.11 | 6.92 | 6.42 | 6.26 | 7.07 |
| O. Wilde              | 7.11 | 7.53 | 7.18 | 7.35 | 8.05 |
| R. L. Stevenson       | 7.2  | 7.71 | 7.11 | 7.16 | 8.06 |
| AVG                   | 6.67 | 7.17 | 6.75 | 6.67 | 7.51 |

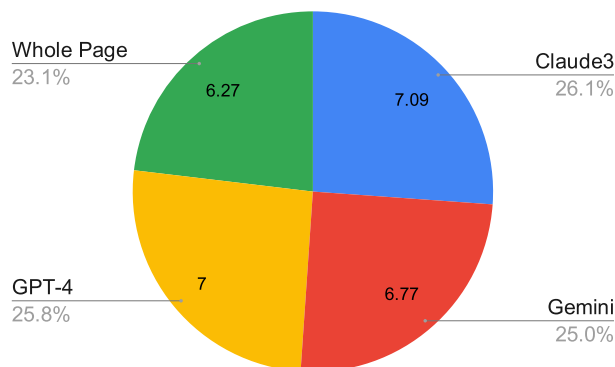
**Fig. 9.** Average image scores between AI language model to book.

#### 4.4. Automation score for generated image results analysis

**CLIPScore.** The diagrams (see Figure 11) contain the averaged data of CLIPScore metrics calculated for the generated images depending on the language model used for each book under test. In general, the results of the metrics can be described as low match: the image does not match the description or the text does not adequately describe the visual content. This result is quite expected, since:

1. This task is quite difficult for GAN models, which have only recently appeared and require a specific input description to generate an image.
2. GAN models, since the models used to calculate the CLIPScore metric were trained on word-phrase-image pairs, which turned out to be incorrect when using abstractions.





**Fig. 10.** Aggregated image ratings across used AI language models.

**Table 6.** Average score between AI language model to book.

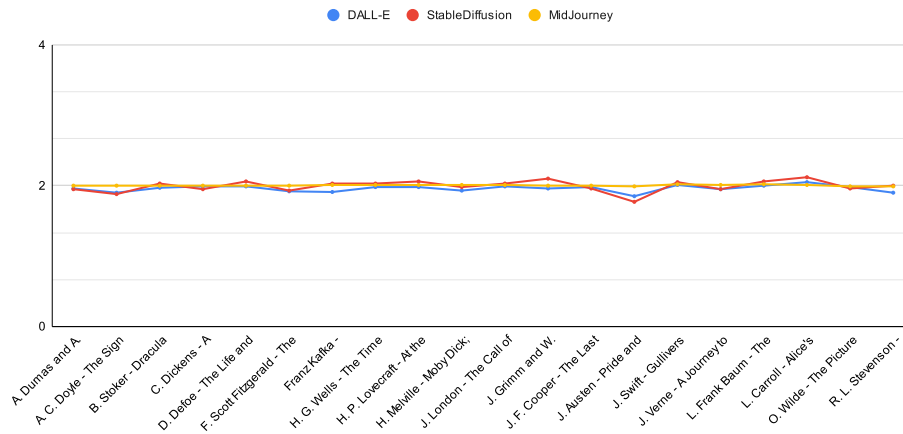
|                       | Claude3 | Gemini | GPT-4 | Whole Page |
|-----------------------|---------|--------|-------|------------|
| A. Dumas              | 6.81    | 6.34   | 6.98  | 5.07       |
| A. C. Doyle           | 7.13    | 6.67   | 7.05  | 6.13       |
| B. Stoker             | 6.85    | 6.61   | 6.61  | 6.38       |
| C. Dickens            | 7.39    | 7.33   | 7.19  | 6.27       |
| D. Defoe              | 7.29    | 7.07   | 7.17  | 5.8        |
| F. S. Fitzgerald      | 7.17    | 6.71   | 6.8   | 6.53       |
| F. Kafka              | 7.25    | 6.71   | 6.6   | 6.4        |
| H. G. Wells           | 6.88    | 6.76   | 6.97  | 5.87       |
| H. P. Lovecraft       | 6.4     | 6.02   | 6.53  | 5.63       |
| H. Melville           | 7.09    | 7.17   | 7.28  | 6.6        |
| J. London             | 7.09    | 6.13   | 6.8   | 6.35       |
| J. Grimm and W. Grimm | 7.05    | 6.52   | 6.98  | 7.07       |
| J. F. Cooper          | 7.53    | 7.23   | 7.4   | 6.47       |
| J. Austen             | 7.72    | 7.48   | 7.79  | 7.43       |
| J. Swift              | 6.04    | 5.54   | 5.61  | 5.32       |
| J. Verne              | 7.05    | 6.95   | 7.17  | 6.33       |
| L. F. Baum            | 7.23    | 7.02   | 7.14  | 6.55       |
| L. Carroll            | 6.91    | 6.6    | 6.77  | 6.27       |
| O. Wilde              | 7.45    | 7.21   | 7.67  | 6.47       |
| R. L. Stevenson       | 7.56    | 7.29   | 7.49  | 6.53       |
| AVG                   | 7.09    | 6.77   | 7     | 6.27       |

All the results obtained are in the same range with minimal difference, which makes the assessment of the highest or lowest indicators irrelevant.

The summarized averages of CLIPScore metrics for images generated by GAN models are shown in table (see Table 7).

**Inception score.** The diagrams (see Figure 12) contain the average data of the Inception Score metrics calculated for the generated images depending on the language model used for each book under test, comparing the abstractions obtained from the AI language models. These results are very similar to the results of the ClipScore metrics and can be described as low match: the image does not match the description or the text does not adequately describe the visual content. Such metrics results are also quite expected and can be explained by the same reasons as for the CLIPScore results.

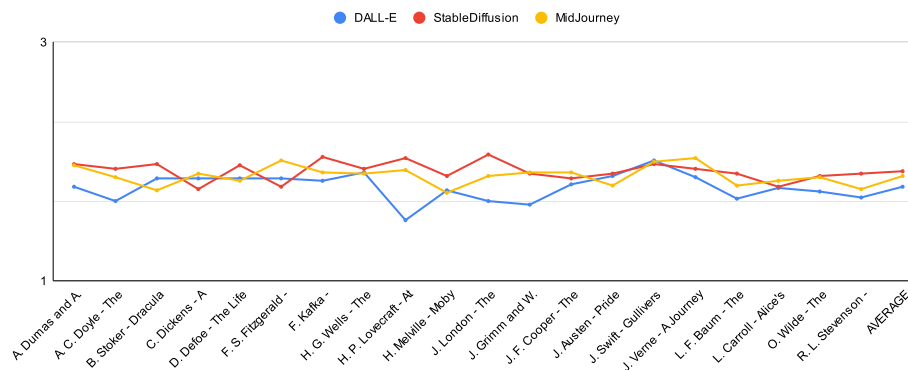
All the results are in the same range with a minimal difference, which makes the assessment of the highest or lowest scores irrelevant.



**Fig. 11.** Average image CLIPScore for GAN models to book.

**Table 7.** Average CLIPScore for GAN models.

|                       | DALL-E | StableDiffusion | MidJourney |
|-----------------------|--------|-----------------|------------|
| A. Dumas              | 1.96   | 1.95            | 2          |
| A. C. Doyle           | 1.9    | 1.88            | 2          |
| B. Stoker             | 1.97   | 2.03            | 2          |
| C. Dickens            | 1.99   | 1.95            | 2          |
| D. Defoe              | 1.99   | 2.06            | 2          |
| F. S. Fitzgerald      | 1.92   | 1.93            | 2          |
| F. Kafka              | 1.91   | 2.03            | 2.01       |
| H. G. Wells           | 1.98   | 2.03            | 2.01       |
| H. P. Lovecraft       | 1.98   | 2.06            | 2.01       |
| H. Melville           | 1.93   | 1.98            | 2.01       |
| J. London             | 1.99   | 2.03            | 2.01       |
| J. Grimm and W. Grimm | 1.96   | 2.1             | 2          |
| J. F. Cooper          | 1.98   | 1.96            | 2          |
| J. Austen             | 1.85   | 1.77            | 1.99       |
| J. Swift              | 2.01   | 2.05            | 2.02       |
| J. Verne              | 1.95   | 1.95            | 2.01       |
| L. F. Baum            | 2      | 2.06            | 2.02       |
| L. Carroll            | 2.05   | 2.12            | 2.01       |
| O. Wilde              | 1.98   | 1.96            | 1.99       |
| R. L. Stevenson       | 1.9    | 2               | 1.99       |
| AVG                   | 1.96   | 2               | 2          |



**Fig. 12.** Average image Inception Score for GAN models to book.

**Table 8.** Average Inception Score for GAN models.

|                       | DALL-E | StableDiffusion | MidJourney |
|-----------------------|--------|-----------------|------------|
| A. Dumas              | 1.79   | 1.98            | 1.97       |
| A. C. Doyle           | 1.67   | 1.94            | 1.87       |
| B. Stoker             | 1.86   | 1.98            | 1.76       |
| C. Dickens            | 1.86   | 1.77            | 1.90       |
| D. Defoe              | 1.86   | 1.97            | 1.84       |
| F. S. Fitzgerald      | 1.86   | 1.79            | 2.01       |
| F. Kafka              | 1.84   | 2.04            | 1.91       |
| H. G. Wells           | 1.91   | 1.94            | 1.90       |
| H. P. Lovecraft       | 1.51   | 2.03            | 1.93       |
| H. Melville           | 1.76   | 1.88            | 1.74       |
| J. London             | 1.67   | 2.06            | 1.88       |
| J. Grimm and W. Grimm | 1.64   | 1.9             | 1.91       |
| J. F. Cooper          | 1.81   | 1.86            | 1.91       |
| J. Austen             | 1.88   | 1.9             | 1.8        |
| J. Swift              | 2.01   | 1.98            | 2          |
| J. Verne              | 1.87   | 1.94            | 2.03       |
| L. F. Baum            | 1.69   | 1.9             | 1.8        |
| L. Carroll            | 1.78   | 1.79            | 1.84       |
| O. Wilde              | 1.75   | 1.88            | 1.87       |
| R. L. Stevenson       | 1.7    | 1.9             | 1.77       |
| AVG                   | 1.79   | 1.92            | 1.88       |

## 5. Conclusions

After analyzing the results, it can be concluded that the described method and the used samples, together with AI language models, cope with the task, although not perfectly. The generated annotations often contain too general a description of the main essence of the text, and it is difficult for GAN models to focus on specific characters or events to display the overall picture.

In general, the results for all GAN models are not very high: for the purpose of the study, the images should have corresponded to scores in the range of 9–10 points to ensure that the images correspond to the text and are realistic; in turn, the data obtained is 27.9–37.3% lower than planned. As a result, it can be concluded that GAN models at the current stage of development cannot provide sufficient quality and cannot generate images that are 100% consistent with the input query. In particular, among the GAN models, the highest results were obtained for the MidJourney model – 7.21 points. The DALL-E model obtained quite similar results, which are 0.69% lower than MidJourney. The StableDiffusion model, in turn, received the lowest scores among the tested models, which is 11.7% lower than the results of MidJourney. Therefore, since DALL-E and MidJourney received the highest scores among the models under consideration, it will be appropriate to continue working with them, looking for ways to improve the results.

The results of the comparison of the used prompts show that prompt P5 received the highest score – 7.51 points. This prompt does not specifically describe how to obtain scene details or characterization, but rather instructs the language model to generate such an abstraction so that the GAN model can use it to generate the best and most appropriate image. This result may indicate that the language models have some insight or knowledge of how GAN models work, which in turn improves the quality of the generated images. Prompt P2 demonstrated the second best result, which is 4.52% lower, but relatively close to the result of P5, and it gives a command to obtain a more general description by selecting only one scene from the text. The rest of the prompts demonstrated results 10.11–11.18% lower, which indicates their limited feasibility for further practical application.

Comparing the results of the language AI models for the generated abstractions and images generated using only the text from the work's page, it can be concluded that the used models perform relatively well in this task, with Claude3 showing the highest results, GPT-4 1.2% lower, and Gemini 4.5% lower. The worst average score, 11.5% lower than Claude3, was obtained when using only page text, which demonstrates that this approach is not relevant and cannot provide sufficient quality of the generated images.

The obtained data of the CLIPScore and Inception Score metrics show that it is not currently relevant to use them for automatic evaluation of generated images, since the results obtained are not correct and differ significantly from the evaluation of the same images by humans. The use of word-image pairs for training GAN models does not provide a sufficient level of quality of the generated images, including when using neural networks to calculate CLIPScore and Inception Score metrics. To obtain results of proper, realistic quality, GAN models should be able to work with descriptive text, and not just search for possible word pairs on which the model was trained.

The limitations associated with training GAN models on keyword-image pairs reduce the practical value of this approach, even if it achieves the best results in generating abstractions. To further improve the method, it is necessary to train models on more generalized concepts. At the same time, the development and training of a separate model exclusively for this purpose is impractical, as it requires significant amounts of data, which are often protected by copyright, or requires obtaining such data in ways that contradict the very idea of automated image generation.

It is worth noting that the assessment of the appropriateness or similarity of the chosen words, abstractions, and generated images is quite subjective, as any reader can imagine and perceive the text in their own subjective way, but the task of this study is not to guess absolutely accurately and convey the formed idea of a certain page before the reader imagines it, but to help imagine, suggest, or form a certain direction of images and objects to better shape the perception of the work.

- 
- [1] Niland A. Picture Books and Young Learners' Reading Identities. *Read Teach.* **74** (5), 649–654 (2021).
  - [2] Niland A. Picture Books, Imagination and Play: Pathways to Positive Reading Identities for Young Children. *Education Sciences.* **13** (5), 511 (2023).
  - [3] Ghazanfari M., Ziaee M., Sharifianfar E. The Impact of Illustrations on Recall of Short Stories. *Procedia – Social and Behavioral Sciences.* **98**, 572–579 (2014).
  - [4] O'Keefe E. J., Solman R. T. The Influence of Illustrations on Children's Comprehension of Written Stories. *Journal of Reading Behavior.* **19** (4), 353–377 (1987).
  - [5] Willett A. The Effect of Text Illustrations on Young Children's Vocabulary Acquisition and Construction of Meaning During Storybook Read Alouds. (2006).
  - [6] Wasylenko K., Tapajna N. The effect of positive and negative illustrations on text recall. *University of Ottawa.* 105–112 (2001).
  - [7] Vavra K. L., Janjic-Watrich V., Loerke K., Phillips L. M., Norris S. P., Macnab J. Visualization in science education. *Alberta Science Education Journal.* **41** (1), 22–30 (2011).
  - [8] Hockley W. E. The picture superiority effect in associative recognition. *Memory & Cognition.* **36**, 1351–1359 (2008).
  - [9] Carney R. N., Levin J. R. Pictorial Illustrations Still Improve Students' Learning from Text. *Educational Psychology Review.* **14**, 5–26 (2002).
  - [10] How This AI Image Won a Major Photography Competition. <https://www.scientificamerican.com/article/how-my-ai-image-won-a-major-photography-competition/>.
  - [11] AI image wins top prize in photography contest...again. <https://www.diyphotography.net/ai-image-wins-top-prize-in-photography-contest-again/>.
  - [12] Lu Z., Huang D., Bai L., Liu X., Qu J., Ouyang W. Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images. Preprint arXiv:2304.13023 (2023).
  - [13] Haase J., Djurica D., Mendling. The Art of Inspiring Creativity: Exploring the Unique Impact of AI-generated Images. *AMCIS 2023 Proceedings.* 10 (2023).

- [14] Aktay S. The usability of Images Generated by Artificial Intelligence (AI) in Education. *International Journal of Technology in Education*. **6** (2), 51–62 (2022).
- [15] Lee Y. H., Chiu C. Y. The Impact of AI Text-to-Image Generator on Product Styling Design. *HCI 2023. Lecture Notes in Computer Science*. **14015** (2023).
- [16] Mookherjee S., Dutta S., Maschatak D., Chakraborty S., Sinha A. Comparative Analysis of Pretrained Text to Image Models for Accurate Radiological Image Generation for a Single Text Prompt. *JMIR Preprints*. 20/07/2023:51099 (2023).
- [17] Abdallah Y., Estévez A. Biomaterials Research-Driven Design Visualized by AI Text-Prompt-Generated Images. *Design*. **7** (2), 48 (2023).
- [18] Göring S., Ramachandra Rao R. R., Merten R., Raake A. Analysis of Appeal for Realistic AI-Generated Photos. *IEEE Access*. **11**, 38999–39012 (2023).
- [19] Zecca P. A., Reguzzoni M., Brambilla A., Protasoni M., Borgese M., Raspanti M. The Dark Side of Artificial Intelligence: The Possible Risk of Falsifying Images for Scientific Articles. *Microscopy and Microanalysis*. **29** (5), 1688–1693 (2023).
- [20] Sarhan H., Hegelich S. Understanding and evaluating harms of AI-generated image captions in political images. *Frontiers in Political Science*. **5**, 1245684 (2023).
- [21] Images: Evaluating Images. <https://guides.lib.uw.edu/c.php?g=344258&p=2318783>.
- [22] Salimans T., Goodfellow I., Zaremba W., Cheung V., Radford A., Chen X. Improved Techniques for Training GANs. Preprint arXiv:1606.03498 (2016).
- [23] Barratt S., Sharma R. A Note on the Inception Score. Preprint arXiv:1801.01973 (2018).
- [24] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Preprint arXiv:1706.08500 (2017).
- [25] Zhang R., Isola P., Efros A., Shechtman E., Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Preprint arXiv:1801.03924 (2018).
- [26] Wang Z., Bovik A., Sheikh H., Simoncelli E. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions*. **13** (4), 600–612 (2004).
- [27] Wang Z., Simoncelli E., Bovik A. Multiscale structural similarity for image quality assessment. *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. **2**, 1398–1402 (2003).
- [28] Erfurt J., Helmrich C. R., Bosse S., Schwarz H., Marpe D., Wiegand T. A Study of the Perceptually Weighted Peak Signal-To-Noise Ratio (WPSNR) for Image Compression. *2019 IEEE International Conference on Image Processing (ICIP)*. 2339–2343 (2019).
- [29] Kynkäänniemi T., Karras T., Laine S., Lehtinen J., Aila T. Improved Precision and Recall Metric for Assessing Generative Models. Preprint arXiv:1904.06991 (2019).
- [30] Hessel J., Holtzman A., Forbes M., Bras R., Yejin C. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Preprint arXiv:2104.08718 (2021).
- [31] Singh J., Zheng L. Divide, Evaluate, and Refine: Evaluating and Improving Text-to-Image Alignment with Iterative VQA Feedback. Preprint arXiv:2307.04749 (2023).
- [32] Yakymiv V. S., Piskozub Y. Z. Research on the use of AI for Selecting Abstractions for Natural Language Image Generation Tools. *International Journal of Computing*. **23** (4), 637–654 (2024).
- [33] Yakymiv V. S., Piskozub Y. Z. Comparison of the use of AI services based on general natural language for generating images for fiction. *Mathematical Modeling and Computing*. **12** (1), 283–298 (2025).

## Порівняльний аналіз використання інструкцій для мовних моделей та автоматизованих метрик оцінювання якості зображень, згенерованих GAN-моделями

Якимів В. С.<sup>1</sup>, Піскозуб Й. З.<sup>1,2</sup>, Оліярник Н. Р.<sup>1</sup>

<sup>1</sup>Кафедра прикладної математики, Національний університет “Львівська політехніка”,  
вул. Степана Бандери, 12, 79013, Львів, Україна

<sup>2</sup>Кафедра прикладної математики, Краківська політехніка,  
вул. Варшавська, 24, 31-155, Краків, Польща

У роботі досліджено можливості застосування мовних AI моделей у поєднанні з генеративними нейронними мережами (GAN) для створення зображень за текстовими описами з художніх творів. Проаналізовано ефективність різних типів промптів, які використовуються для формування абстракцій, а також проведено порівняння результатів роботи найпопулярніших сучасних моделей генерації зображень – MidJourney, DALL-E та Stable Diffusion. Отримані результати демонструють, що хоча мовні моделі здатні формувати змістовні абстракції, які частково відповідають текстовому змісту, самі GAN моделі на сучасному етапі розвитку не забезпечують необхідного рівня відповідності та реалістичності. MidJourney показала найвищі результати, в той час як DALL-E поступилася всього на 0.69%, а StableDiffusion на 11.7%. Результати оцінювання показали перевагу промптів, що покладаються на генерацію узагальнених абстракцій, а промпт який повністю покладається на мовну модель у питанні генерації абстракції перевершує інші промпти на 4.52–11.18%. Натомість автоматизовані метрики як-от CLIPScore та Inception Score виявилися некоректними для цієї задачі. Також висвітлено обмеження сучасних підходів до навчання GAN моделей на парах “ключове слово — зображення” та обґрунтовано потребу в удосконаленні методів генерації на основі повноцінних текстових описів.

**Ключові слова:** штучний інтелект; обчислення; ai-зображення; генерація зображень; перетворення тексту в зображення.