

# New Frontiers in Metaphor Visualization: An Approach Through Generative Artificial Intelligence<sup>1</sup>

Metaphor Visualization Through Generative AI

Original Study

Cerutti Margaux; LUMSA University; m.cerutti.dottorati@lumsa.it  
[Lis Mateusz Mirosław; SophIA; mateuszmiroslawlis@sophia.vision](mailto:Lis.Mateusz.Miroslaw.SophIA@mateuszmiroslawlis@sophia.vision)  
Polidoro Piero; LUMSA University; p.polidoro@lumsa.it

Received: 29 May 2024; Accepted: 13 July 2024

**Abstract:** In the last two years, our culture and our everyday lives have literally been invaded by the use of new technological systems based on the operation of artificial intelligence. In a very short time, we have seen the development of new platforms capable of generating images or texts autonomously, following only instructions (input) of a few words. From this thought comes our question. What would happen if we asked such a system to produce something using a metaphor as input? The metaphor was chosen because, according to important semiotic and linguistic theories (Eco, Lakoff & Johnson), it is one of the foundational elements of human thought. As we will see, generative artificial intelligence systems exhibit different behaviors when faced with metaphorical inputs. In this article, we will attempt to explore these behaviors and the variables that trigger them through various approaches.

**Keywords:** Artificial Intelligence; Generative AI; Metaphor; Prompt Engineering; Semiotics

## 1. ARTIFICIAL INTELLIGENCE. FROM PREDICTING TO GENERATING

When faced with a new technology, it is always interesting to test its limits, push it to the boundaries of its scope, and understand how it reacts. This is even more true with a tool like generative artificial intelligence, which is seen (rightly or wrongly) as “opaque”, i.e., inscrutable in its internal workings. Given the great achievements that verbal generative artificial intelligence systems, based on LLM, are achieving and given that, as we will see, for many scholars the metaphor deeply permeates our linguistic

and, consequently, conceptual systems, the temptation to test generative AIs with non-literal expressions, but figurative and, specifically, metaphorical, is great. This is what we have done in some attempts, aimed at both LLM and TTI systems. Before exploring the systems just mentioned, it is necessary to account for some technical innovations that have allowed the transition from predictive AI to generative AI.

The increase in web users and the massive production of data have made available huge amounts of images that can be used for the training of image classification

<sup>1</sup> This research is part of the activities carried out within the framework of the European Union HORIZON-RIA SOLARIS project, grant number 101094665. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

This chapter is a collaborative work of the three authors. However, the writing of the text can be attributed to Margaux Cerutti for paragraph 1, Piero Polidoro for paragraphs 2 and 3, and Mateusz Mirosław Lis for paragraph 4; paragraph 5 was written by all three authors.

and recognition systems (Chen et al. 2020). Thanks to statistical models, neural networks are trained, increasing their computing capacity, and making them very versatile. In 2006, Geoffrey Hinton introduced Deep Learning (DL), explaining how these new algorithms can be trained to recognize objects in images and videos autonomously (LeCun et al. 2015). The development of Transformers was another milestone. Transformers are DL algorithmic architectures more specialized in the processing and interpretation of sequential data in parallel (Fernández Gamìn et al. 2024). Transformers are a new system that includes the encoding of the context of words within a text. Although the introduction of DL has brought decisive improvements it is still a predictive AI. Regarding the introduction of generative AIs, it is important to dwell—briefly—on Natural Language Processing (NLP) and Computer Vision that have led to the creation and diffusion of Large Language Models (LLM) and Text-To-Image (TTI) models.

### 1.1 LLM SYSTEMS

Natural Language Processing (NLP) is a technique of interaction through natural language that brings together linguistic and computer approaches since the 1950s. These are algorithms capable of analyzing and using natural language. Initially, all models were based on deterministic systems; only from the 1980s the symbolic-deterministic approach was abandoned favoring statistical systems, which take into consideration the frequency of a word, the average length of sentences and so on. This choice has allowed the adaptation and training of models that, however, not yet having access to large datasets, were still very limited. Subsequently, thanks to the diffusion of neural networks, a new way of representing and processing the words of textual corpora was introduced in 2004: the Word Embeddings. This has allowed a new processing of natural language mapping words or phrases into vectors of real numbers. In other words, the vectors of semantically similar words are found closer within a defined space (Bernard 2024), i.e., a similar vector representation, allowing the storage of both semantic and syntactic information. For example, the words “woman” and “lady” will be close in the vector space. Subsequent to the introduction of Transformers, Jacob Devlin presented in 2018 BERT (Bidirectional Encoder Representations from Transformers). It is an NLP model based on automatic learning that has represented a significant improvement in text understanding and in performing a variety of tasks. In the same year, OpenAI introduced its first GPT model: Generative (able to generate an output), Pre-trained (trained on a dataset of texts) and Transformer (i.e., based on transformer technology). These models have allowed the development of the LLMs used for generative AI, such as ChatGPT.

### 2.2 TTI SYSTEMS

Computer Vision was born in the 1960s. The first Image recognition systems were developed, as well as the first predictive AI models, through the adoption of various approaches. It is necessary to keep in mind, however, that until the end of the 1990s, these systems were mainly theorized or applied in very specific fields. For example, Image Segmentation allowed to group (clustering) the pixels in an image. In other words, the machine could be asked to group only the pixels belonging to the image of a person, separating them from the background or any other objects (Kaur, D., Kaur, Y. 2014).

From the 2000s, researchers’ attention has focused on object recognition and in particular on faces. A crucial role for the development of these algorithms was the introduction of digital cameras, which led to a decisive increase in the quality and quantity of available images. This allowed the creation of huge datasets, such as ImageNet (2009), fundamental for the training of detection and recognition algorithms with DL models .

Up to that point, it was still predictive AI. We would have to wait until 2014 to be able to talk about generative AI, when Ian Goodfellow introduced the GANs, Generative Adversarial Networks capable of generating original images through automatic training on very large datasets (Fernández Gambín et al., 2024), definitively opening the doors to generative AI for images. From this moment, research has focused on the development of models aimed at increasingly improving the recognition, classification and generation of images. In particular, also in 2014, the Denoising Diffusion Probabilistic Models (DDPM), also known as diffusion models, are presented. They have allowed one to easily achieve the generation of images of exceptional quality in a very short time (Ho et al. 2020). Subsequently, in 2021, the CLIP model is introduced, which works with both text and images and will be discussed in depth in the last part of this paper. These innovations led in 2021 to the presentation of DALL-E, the first TTI model capable of generating images using both the CLIP model and a GPT model. Finally, in 2022 VQGAN-CLIP is also presented, another TTI model based on the combination of CLIP and GAN. Today, however, most systems – such as Stable Diffusion and DALL-E 2 – use diffusion models and CLIP.

For the purposes of this contribution, we will proceed with the presentation of some of the main TTI systems. However, it must be noted that most of the systems we will present have been trained on datasets like LAION<sup>2</sup>—the only one that stands out in this case is, as we will see, Firefly. LAION is a non-profit organization that provides datasets and models for free access for research in the field of automatic training, to promote public education and a more ecological use of resources. In 2022, LAION-5B<sup>3</sup>, a large-scale research dataset, was presented by Romain Beaumont. It is the largest multimodal dataset

2 Link: <https://laion.ai/> - Accessed on 26/05/2024.

3 Link: <https://laion.ai/blog/laion-5b/> - Accessed on 26/05/2024.

in the world, composed of several billion image-text pairs filtered through CLIP. Despite being freely accessible, it is an unsupervised dataset, so it has not been carefully selected, filtered or modified to remove unwanted or potentially harmful content. In other words, it has not undergone a review process.

*Midjourney*, launched in 2022 by David Holz. It does not work through a website but is run through Discord<sup>4</sup>. It is not open source<sup>5</sup>, it only has an online function based on the cloud and all the images generated are distributed with a Creative Commons license<sup>6</sup>.

*DALL-E*, launched in 2021 by OpenAI<sup>7</sup>. It was one of the first software to become known worldwide. Also in this case it is based on the cloud, it only has the online function and it is not open source. Unlike *Midjourney*, however, *DALL-E* operates with greater transparency, through the publication of several articles that explain the stages of creation and its operation. *DALL-E* is certainly recognized for the high quality of the images generated in a very short time and for the versatility in the interpretation of verbal text – being also associated with a GPT model. The big disadvantage concerns some restrictions in the generation of different elements and high costs in case of intensive use. The restrictions in the generation suggest that, although the system was trained with an unsupervised dataset, subsequent modifications have been made that allow greater control of the generations. The transparency in the communication operated by OpenAI ensures the understanding of the functioning of the system; for example, we know for sure that also the subsequent versions, i.e. *DALL-E 2* and *3*, use diffusion models.

*Stable Diffusion*, published in 2022 is the only completely free system, available both online and offline – with appropriate setup – and open source. However, these advantages and this great transparency have not made the platform accessible to everyone. In fact, it requires great technical skills and adequate hardware resources. Also in this case the training took place on LAION-5B, which as we have seen, being an unsupervised dataset can contain a wide range of unverified information.

*Firefly* is the generative AI model presented by Adobe. It is a very simple web application and has a great advantage: unlike all the applications presented so far, *Firefly* was trained using a licensed dataset (created from the Adobe Stock service) and is therefore compatible with commercialization<sup>8</sup>. This advantage, however, brings with it a disadvantage, which concerns the size of the dataset, which is relatively small and does not include, for example, what is defined as elements of *pop culture* (commercial brands, etc.).

The use of unsupervised datasets can have various consequences; in particular they can contain distorted representations of reality. For example, they can be characterized by over-representations of some categories or groups at the expense of others. This happens because the datasets used for training are generally created from the data we can find on the web. To retrieve this data, the technique of *scraping* is used, using software similar to those of the most common search engines to label online content. Training on this type of dataset involves the transfer of biases that already characterize the content scraped from the web. The various attempts made by Google over time, demonstrate how these databases amplify certain prejudices and discriminations present in society. For example, in 2015 it was discovered that the Word2Vec program produced distorted results, such as the refusal to classify women as doctors (Garrido-Muñoz et al. 2021). In the same year, it was shown how some Google systems erroneously labelled as “gorilla” the images of people with dark skin. To remedy this error, Google simply removed the label, which still in 2023 was not reintegrated into the database. These cases presented are obviously extreme cases, but if we compare the outputs generated by different systems, we quickly realize that some characteristics – albeit not specified – are present in all images. For example, if we ask any TTI to generate the image of a dog, almost always the image will represent a dog with sweet and joyful features and soft fur.

## 2. SEMIOTIC MODELS FOR VERBAL AND VISUAL METAPHORS

The metaphor certainly represents one of the central and most discussed themes of rhetoric. Aristotle already paid great attention to it, even if more in the *Poetics* than in the *Rhetoric*. In fact, Aristotelian rhetoric is mainly aimed at argumentation: *inventio* is its main axis, while *elocutio* (style issues) are only marginal. It is instead in literature (and in particular in tragedy) that style acquires greater importance, with all its deviations and its ornaments (including the metaphor).

In the Twentieth century, after a period of discredit, rhetoric again attracted the attention of scholars. In the semiotic field, between the 1960s and the 1990s, the Group  $\mu$  (first in the *Retorica generale* and then in the *Traité du signe visuel*) and Umberto Eco (1984) were particularly interested. For the Group  $\mu$  the rhetorical mechanisms are fundamentally based on the fact that there is a *deviation* from a norm: this deviation pushes the user (who will have to interpret the metaphor) to linger on the text, to ask questions and to look for connections and further meanings. In the specific case of tropes (i.e.,

4 Free messaging and distribution platform.

5 When a system is not open source, it means that it has not been programmed to be publicly accessible, so it is not known on which dataset the training takes place.

6 In other words, they can be used by any other user.

7 OpenAI is an AI lab that promotes the development and improvement of new technologies.

8 The intellectual properties of images generated by generative AI are currently the subject of debate that we will not discuss.

substitutions that occur at the level of single words; the metaphor is an example) we therefore have a deviation from what is the literal meaning that we expect. If I say that “a maiden is a reed”, this sentence makes no sense from a literal point of view. But from a figurative point of view, it means that the maiden (the metaphorized) is slender and flexible like a reed (the metaphorizer). In the perspective of the Group  $\mu$  the metaphor is nothing more than the concatenation of two synecdoches, one generalizing and the other particularizing (but the order does not necessarily have to be this). In the case of the maiden, in fact, to understand the metaphor I must first proceed by generalization and reduce the species “maiden” to the genus “slender and flexible things”; then I can descend to another species belonging to this genus: the reed. The genus in this case acts as a bridge between the two terms (the metaphorized and the metaphorizer) and highlights how between them there is a similarity (being slender, flexible).

In his discussion on the metaphor Eco (1984) starts from the Group  $\mu$ , criticizing its thesis. Eco, in fact, has already developed a semantic theory that abandons dictionary structures (the so-called Porphyrian trees, derived from an Aristotelian setting) and instead embraces “encyclopedic” ones, more akin to cognitive sciences. In encyclopedic semantics, there are no pre-established hierarchical structures, but nuclei of meaning (sememes) that are composed of semantic elements, called semes. Depending on the context and use, sememes and semes can exchange roles (so a seme can become a new sememe to decompose, and vice versa) and reorganize themselves into new networks of relationships. Once this is clarified, it is understood how for Eco the metaphor is a substitution between two sememes that are united by one or more semes, which will be the intermediaries of the semantic exchange between metaphorizer and metaphorized. From here, Eco comes to propose five rules to guide the metaphorical interpretation.

A very important aspect of Eco’s discussion is his attention to the cognitive value of the metaphor (which was already present in Aristotle). The metaphor, that is, not only serves to amaze or embellish but also makes us know new aspects of things and stimulates our reflection on oppositions and analogies between cultural units. This is the same approach followed by two important American scholars, George Lakoff and Mark Johnson, in their classic work *Metaphors We Live By* (1980). The idea of Lakoff and Johnson is to study this conceptual system (the way concepts are structured within them and between them) through its manifestation in verbal language.

Studying language Lakoff and Johnson realize something very interesting: our speeches are woven with metaphors, much more than one might think. These are not, however, the “poetic” metaphors, those that, easily recognizable, enrich a text and develop an aesthetic function (metaphors such as “Laura has golden hair”, to understand). They are, instead, much more subtle

metaphors, sometimes almost unrecognizable, but that we use continuously.

Let’s think, for example, of expressions like “he attacked a weak point in my argument”, “his criticisms hit the mark”, and “I’ve never *won* an argument with him”. They are all governed by a basic metaphor: ARGUMENT IS WAR.

The hypothesis of Lakoff and Johnson is that the conceptual system is often organized metaphorically. Some concepts (for example, that of argument) are indeed structured through metaphors based on other concepts (in our case, war). This can happen, for example, with those concepts that would otherwise be unclear, because they are particularly abstract or far from our experiences. Their metaphorization then helps us to see them in terms of another concept that is more familiar to us and therefore allows us to understand them more. The consequence, however, is that their meaning will be transformed, or at least bent. Metaphorizing the argument through the concept of war means that, from then on, we will see the argument as a war and therefore, substantially, in conflictual terms. At the beginning of Chapter 1 of their book, Lakoff and Johnson wonder what would happen, in practical terms, in a culture where argument is seen not as a war or a fight, but as a more cooperative activity, such as a dance.

To what has been said so far, we must add an important piece, which concerns the visual. Rhetorical mechanisms and the metaphorical one in particular do not only concern verbal language. We can, for example, speak of metaphor also in the field of visual language, both because, as we will see, there are images that are based on a metaphorical mechanism, and because the common seeds that are at the base of the metaphorical exchange can be of various nature.

How to describe and classify visual metaphors? This is not the place for an in-depth discussion (for which we refer to Sonesson 1989 and Group  $\mu$  1992), but we can provide some basic elements. First of all, we must distinguish the metaphors *in absentia*, i.e., those in which the metaphorizer replaces the metaphorized (which is, precisely, absent) and those *in praesentia*, in which instead the two elements coexist within the image (it would be equivalent to that specific type of metaphor that is the simile). According to the Group  $\mu$ , however, this opposition is not sufficient. It is necessary to consider that visual elements can recombine and hybridize (as in a linguistic crisis) in order to occupy the same space. So, the opposition *in absentia/in praesentia* will intersect with that between *conjunct* elements (i.e., that occupy or would have occupied the same position) and *disjoint* elements, giving rise to four possible situations.

### 3. METAPHORS AND VISUAL REPRESENTATION THROUGH TTIS: A SEMIOTIC ANALYSIS

We asked ourselves, what happens if we provide a TT1 system with a prompt containing metaphors. We tried, for example, to insert in Midjourney this prompt<sup>9</sup>: Achilles strong as a lion

9 The tests with Midjourney were carried out in July 2023. Those, which follow, with DALL-E in May 2024.



Figure 1. Midjourney's reply to the prompt "Achilles strong as a lion".

The result is four images (which we number from 1 to 4, clockwise, starting from the one in the upper left); we do not refine the prompt further, because we would risk providing our own metaphorical interpretation. In these images we can notice some things. In three out of four cases (1, 3, 4) the physiognomy of Achilles seems inspired more than by a classical physiognomy by the appearance of Brad Pitt, the actor who impersonated him in *Troy*, a famous Hollywood film of 2004. It should also be said that using the word "Achilles" to do an image search on Google or other search engines, results representing Brad Pitt in his role as Achilles are not a majority, but certainly are among the most frequent and could therefore have influenced the databases on which the systems were trained.

The second consideration is that in two cases (1, 4) a part of the prompt is practically omitted: there is in fact only Achilles, with a classic scenery in the background.

In two other cases (2, 3), instead, the lion is present in the background (behind Achilles, in 2, or to his left, in 3). The mechanism could therefore seem that of the metaphor *in praesentia disjunct*; this, however, usually relies on a visual similarity, to propose a deeper semantic analogy. In this case, instead, more than a visual similarity there is an evident juxtaposition between two connected elements, as could happen in a heraldic composition. The mechanism, therefore, more than that of the metaphor, is that of metonymy, where the elements exchange meanings not by analogy, but by contiguity. In other words, from a linguistic point of view, it is as if the TTIs interpret the "as" that distinguishes metaphors as the "and" that represents a metonymic process. On DALL-E the results are not very different. The prompt, in this case, was "Draw Achilles strong as a lion". The system produced two images.



Figure 2. DALL-E's reply to the prompt "Draw Achilles strong as a lion"..

The images are followed by this comment: "Here are the depictions of Achilles, showcasing his strength and heroic spirit. If you need any adjustments or additional details, let me know!". It is interesting to note how the verbal comment clearly shows a metaphorical interpretation, while the two images reproduce the metonymic mechanism already encountered in Midjourney. Also, note that the resemblance to Brad Pitt diminishes, but

certainly, physiognomies in line with classical canons are not proposed; in the second image, you can also notice a typical hallucination: a headless horse flying, holding a flag with its legs.

We subsequently tried with a metaphor inspired by the work of Lakoff and Johnson, providing Midjourney with this prompt: An argument as a war  
The result is as follows.



Figure 3. Midjourney's reply to the prompt "An argument as a war".

**New Frontiers in Metaphor Visualization: An Approach Through Generative Artificial Intelligence**

The four images share the same perspective setting. Images 2, 3, 4 also share both the use of a very long field and— from an enunciative point of view— the presence of observers from behind, who perform the function of delegates of the person looking at the image. In all three cases, they seem to observe a war landscape more than participate in the conflict. Image 1, on the other hand, stages an enunciative débrayage (third-person vision) and two actors who challenge each other, always in a scenario of war and destruction. This is the image that comes closest to the representation of an active confrontation and therefore to the metaphorized (the argument). But

it must be said that also 3 and 4 present a double actor and therefore a structure potentially compatible with the polemical confrontation. In any case, whether it is explicit (image 1), potential (images 3 and 4) or relegated to the relationship between observer and observed (image 2), the confrontation (the genre to which the metaphorized belongs) always remains quite separate from the metaphorizer (the war), reaffirming that metonymic mechanism that we also identified with the prompt on Achilles. Also in this case we made a second test with DALL-E. The prompt was equivalent, with the addition of the initial "Draw...". The two images it returned are shown below:



Figure 4. DALL-E's reply to the prompt "Draw an argument as a war"

The verbal comment, on the other hand, is not very significant: “Here are the dramatic depictions of an argument as a war. If you need any modifications or additional elements, feel free to let me know!”

The representation seems to show different aspects (and perhaps more advanced) than the previous ones. The structure is similar to that of image 1, with two opposing actors ready for combat. This time, however, at least in the first image, they are at the head of two ranks of opponents. It is interesting to note how while in the images generated by Midjourney “argument” was reduced to the opposition between actors or observers, while the metaphorizer (i.e., the war) prevailed, here in the foreground remains the figurative system linked to “argument” (pens, written sheets, people iconographically recognizable as intellectuals). The war is present at a deeper level, in the topological opposition of the two sides, in the attitude assumed by the contenders and, above all, in the way of holding (*as if* they were weapons) the writing tools.

#### 4. LLM CAN HELP TTI MODELS UNDERSTAND METAPHORS

Most of the cutting-edge tools for image generation, such as DALL-E (Ramesh et al. 2022) and Stable Diffusion (Rombach et al. 2022), employ architectures usually composed of two distinct systems: (i) a “generative” component, which generates images; (ii) a “validation” component, which guides and conditions the first component towards the generation of images that better match the input prompt. Through the repeated application of the two components, these models can converge quickly towards accurate and high-quality images. While the generative phase of this process is realized by models specific to each system—some like ImageGPT employ the Generative Pre-trained Transformer (Chen et al. 2020), others like Stable Diffusion and DALL-E employ diffusion models—the validation component is almost always a Contrastive Language-Image Pre-Training (CLIP) model, i.e., a neural network trained on a dataset of image-text pairs. This model is the core of the guiding process in image generation. A CLIP model encodes both images and texts into a single common geometric space of “embeddings”—i.e., high-dimensional numerical vectors. This encoding is learned by training the model on an image-text pair dataset. A well-trained CLIP model should encode images and corresponding texts with “similar” embeddings. In other words, by tracing the coordinates of the vectors output from the model, the points representing the image of a dog and the caption “a dog” should be very close to each other. The sheer dimensional magnitude of the vectors handled by CLIP allows it to learn abstract concepts and classifications such as temperature information: words like

“fire”, “flame”, “hot” have associated embeddings close to those of images of bonfires and volcanoes while they are well separated and distant from the embeddings of words like “ice” and “cold”.

As previously mentioned, the “correct” way to transform image and text into a common geometric space is learned by training the model on huge datasets of image-text pairs. These datasets are usually built from data collected on the Internet through web-scraping and are mostly made up of image-caption pairs. The caption is generally a literal description of the image and its contents. This is why metaphorical prompts have strong limits with current image generation systems. Even though research (Bondielli, Passaro 2021; Bustos et al. 2023) has shown that pre-trained CLIP models can be reused for other tasks such as Image Emotion Recognition and Visual Sentiment Analysis, the interpretation of visual metaphors is currently a task beyond their reach. These models are built on descriptions and not on interpretations of images.

A first and simple solution to the interpretative limit of TTI systems based on a “literal” CLIP model is to transform the metaphorical prompt into a descriptive prompt by means of a classic Large Language Model (Chakrabarty et al. 2023). Large-scale text generative models boast much more extensive emerging interpretative capabilities than TTI models (Zoph et al. 2022). This fact can be explained by considering the larger size of their learning datasets and the use of a radically different generative structure. The actual capabilities of LLMs in recognizing and interpreting metaphors have already been the subject of pioneering studies that have identified in the most popular LLMs like GPT-3.5<sup>10</sup> both limits (Tong et al. 2024) and effective solutions through specific prompting techniques such as Chain-of-Thought<sup>11</sup> (CoT) (Prystawski et al. 2023). According to Yakura (2023), moreover, the ability of LLMs to understand metaphors scales with the number of parameters of the model, a trend not observed for the recognition of sarcasm. The contrasting results of these first attempts indicate the need for further studies using other advanced techniques and comparing the results on the main benchmarks introduced by recent literature (Tong et al. 2024). The solution proposed by Chakrabarty (et al. 2023) exploits this interpretative advantage of LLMs to transform metaphorical prompts into descriptive prompts to then provide to a TTI system through a hybrid approach of prompting with CoT or few-shot<sup>12</sup>. Their study proposes the statistical evaluation by three human users of the images generated by the cascade of the InstructGPT3 text model with both DALL-E 2 and Stable Diffusion. The users then evaluated how much the final images were a plausible visual representation of the initial metaphors before they were converted by the LLM. From their analysis, the InstructGPT3 configuration

10 The LLM behind ChatGPT.

11 CoT prompting involves making the model reason by forcing it to explain the logical steps of the inference before providing the final result.

12 The prompting few-shot involves presenting the text model with examples of the resolution of the instruction to be executed.

with DALL-E 2 emerges as the winner. Their system is used for the generation of visual elaborations that maintain the literal representation of the elements present in the metaphors of the initial prompts (the proposed elaboration of “My lawyer is a shark” is a shark dressed as a lawyer) and not for the interpretation of the same (which would instead replace “My lawyer is a shark” with “My lawyer is fierce”). By reworking their system we will demonstrate that the same approach can be proposed for both cases. The prompt used by Chakrabarty (et al. 2023) is the following:

### PROMPT:

*Your task will be to elaborate a metaphor with rich visual details along with the provided objects to be included and implicit meaning. Make sure to include the implicit meaning and the objects to be included in the explanation*

1. Metaphor: My lawyer is a shark.

*Objects to be included: Lawyer, Shark*

*Implicit Meaning: fierce*

*Visual elaboration: A shark in a suit with fierce eyes & a suitcase & a mouth open with pointy teeth.*

*[...]*

6. My bedroom is a pig sty

This entire text forms the prompt for the LLM model; the metaphor that will actually be analyzed is inserted in the last line of the prompt, all the preceding text are numerous examples (the so-called “shots”; in this case we are talking about a few-shot approach) that show the LLM what to do and how to write the output. The final prompt for the TTI model is reported after “Visual elaboration.”.

Using a similar but revisited and optimized approach, we are able to transform a metaphorical prompt into a descriptive prompt whilst avoiding literal interpretations. Chakrabarty et al. use InstructGPT3 with a CoT 5-shot prompt, we use GPT-3.5-TURBO with a CoT 1-shot prompt.

### PROMPT:

*Given the text for a visual metaphor, reason and produce a prompt for a text-to-image model to generate an image that is as faithful as possible to your interpretation of the metaphor. The generated image should not be the literal representation of the metaphorical text, but its correct visual interpretation. Please remember that the text-to-image model generates images by interpreting the prompts literally.*

*Examples:*

*VISUAL METAPHOR: Alice has golden hair.*

*REASONING: The most immediate interpretation is that Alice has blonde hair. Gold is often used to describe bright and shiny blonde hair, evoking an image of light and radiant hair. FINAL PROMPT: A girl with blonde hair.*

*Input:*

*VISUAL METAPHOR: Achilles is a lion.*

### OUTPUT:

*REASONING: The visual metaphor “Achilles is a lion” evokes the image of a powerful, valorous, and dominant*

*individual, akin to the lion representing strength and courage. Based on this interpretation, the final prompt should evoke an image that embodies these characteristics without being literally the figure of a lion.*

**FINAL PROMPT:** *A strong and courageous man, with a regal and proud gaze.*

We can afford to use the same explicit prompting approach (i.e., explaining to the LLM the instruction to be performed) as GPT-3.5-TURBO is also instruction-based, i.e., it is a model resulting from a specific pre-training for the execution of conversational instructions. We specify that the exact approach (at the cost of moving from a 1-shot prompt to a 2-shot prompt) produces the same result with a traditional completion-only LLM (i.e., not instruction-based) like DAVINCI-002. The need to add one or more examples of metaphor elaboration (the “shots”) is explained precisely by the lack of “instruction” learning typical of instruction-based models; DAVINCI-002 not having this additional layer of knowledge requires a few more examples to learn the ability to transform metaphorical prompts. It is unproven whether the 2-shot approach is sufficient to convert all possible metaphorical prompts; we remind you that no statistical test of the proposed methods has been made. Therefore, the evaluation of these techniques on different models and dedicated benchmarks would be needed to estimate their actual degree of success in executing the requested instruction.

### PROMPT:

*VISUAL METAPHOR: Alice has golden hair.*

*REASONING: The most immediate interpretation is that Alice has blonde hair. Gold is often used to describe bright and shiny blonde hair, evoking an image of light and radiant hair. FINAL PROMPT: A girl with blonde hair.*

*VISUAL METAPHOR: Mario and Luigi are like cat and dog.*

*REASONING: The visual metaphor “Mario and Luigi are like cat and dog” suggests that Mario and Luigi have a conflicting or antagonistic relationship. In popular culture, dogs and cats are often seen as animals that don’t get along, thus representing an idea of continuous discord or rivalry.*

**FINAL PROMPT:** *Two boys squabbling.*

*VISUAL METAPHOR: Achilles is a lion.*

### OUTPUT:

*REASONING: The most immediate interpretation is that Achilles is brave and bold. Lions are often portrayed as brave and courageous, and are often used as symbols of strength and power.*

**FINAL PROMPT:** *A brave warrior.*

An approach that does not completely induce the LLM to interpret the metaphor is to use an instruction-based model like GPT-3.5-TURBO without the few-shot approach, i.e., without the examples of interpretation and transformation of a metaphorical prompt into its corresponding descriptive prompt.

### PROMPT:

*Given the visual metaphor “Achilles is a lion”, reason*

and produce a prompt for a text-to-image model to generate an image that is as faithful as possible to your interpretation of the metaphor. The generated image should not be the literal representation of the metaphorical text, but its correct visual interpretation. Please remember that the text-to-image model generates images by interpreting the prompts literally.

The new prompts generated are:

*A heroic ancient Greek warrior standing confidently on a battlefield. The warrior should have an imposing, muscular build and wear traditional Greek armor with intricate details, including a helmet with a prominent crest. Capture his fierce determination and strength in his stance and facial expression. **In the background, depict a majestic lion standing beside him**, emphasizing their shared qualities of courage and power. The scene should be set during battle, with hints of conflict in the distant background, but the focus remains on the warrior and the lion, highlighting their commanding presences.*

*A powerful and valiant warrior standing on a battlefield, exuding immense strength and bravery. The warrior, with intense and determined eyes, is adorned in ancient Greek armor **with a lion emblem prominently displayed on his chest**. His posture is fearless and commanding, as if he is ready to take on any challenge. The background shows a dramatic sky and a chaotic battlefield, which underscores his heroic presence and formidable spirit.*

*A heroic warrior in ancient Greek armor, standing boldly on a battlefield, displaying an aura of power and bravery. **He has a lion's mane hairstyle and wears a cape that subtly resembles a lion's fur**. The background features a setting sun casting a golden light, enhancing his majestic and powerful presence.*

The “instruction-only” approach, while correctly identifying the metaphorical connection to courage and ferocity in the example “Achilles is a lion,” generates prompts that retain the word “lion,” leading TTI models to interpret it literally. This behaviour disappears with few-shot prompting as we have demonstrated previously. Even if our experimentation confirms other results present in literature, it is necessary to validate and quantify these approaches through statistical analysis and qualitative evaluations by human experts.

## 5. CONCLUSIONS

The recognition, interpretation, and visual representation of metaphors by both LLMs and TTI models is an open research domain, with the literature still scarce. This work aimed to contribute by exploring the behaviour of generative AIs when tasked, under various conditions and through different approaches, with visually representing metaphors. As we have seen, an instruction-based approach tends to pre-interpret the metaphor more deeply; in other words, it is as if the generative AI in this case performs a reverse-engineering task, bringing the metaphor back to its original meaning, i.e., the characteristic

of the metaphorized subject that was intended to be highlighted. The non-instruction-based approach, on the other hand, although capable of giving a verbal interpretation of the metaphor, tends to represent it visually in a literal way. However, we have seen that this literal visual interpretation often has more of the structures of metonymy than metaphor, transforming the operator “as” into the operator “and”.

Further studies are certainly needed in this direction, using different approaches, including statistical evaluation of different prompting styles for metaphor interpretation through LLMs, the comparison or the creation of new benchmarks for the evaluation of LLM and TTI systems, as well as more advanced analyses and designs interfacing with and modifying the models themselves. This latter opportunity could see the transformation of the problematic CLIP model from a descriptive embedder to a metaphorical one through learning on a dataset of metaphors and corresponding visual representations; such a model would have various direct uses in the graphic, advertising, and audiovisual domains precisely for its ability to make evaluations not necessarily bound to literal composition. Another development opportunity would see the use of the same metaphor dataset for training the LoRA model, which should become capable of inducing systems like Stable Diffusion to a different graphical representation of visual metaphors; as in the previous case, such a system would have clear uses both academically and commercially.

## BIBLIOGRAPHY

- Bernard, J. 2024 . What are word embeddings? IBM, available at: <https://www.ibm.com/topics/word-embeddings>
- Bondielli, A., Passaro, L. C., 2021. Sfruttare CLIP per il riconoscimento delle emozioni nelle immagini. In Cabrio, E., Croce, D., Passaro, L. C., Sprugnoli, R. (Eds.), Atti del Quinto Workshop su Linguaggio Naturale per l'Intelligenza Artificiale (NL4AI 2021) in concomitanza con la 20a Conferenza Internazionale dell'Associazione Italiana per l'Intelligenza Artificiale (AI\*IA 2021), Evento online, 29 novembre 2021. CEUR Workshop Proceedings, 3015. CEUR-WS.org. available at: <https://ceur-ws.org/Vol-3015/paper172.pdf>
- Bustos, C., Civit, C., Du, B. et al., 2023. Sull'uso dei modelli Vision-Language per l'analisi del sentimento visuale: uno studio su CLIP. In 11a Conferenza Internazionale sull'Interazione Intelligente e l'Informatica Affettiva (ACII) (pp. 1-8). IEEE Computer Society. available at: <https://doi.ieeecomputersociety.org/10.1109/ACII59096.2023.10388075>
- Chakrabarty, T., Saakyan, A., Winn, O. et al., 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.
- Chen, M., Radford, A., Child, R. et al., 2020 . Preallenamento generativo dai pixel. In III, H.

- Daumé & Singh, A. (Eds.), Atti della 37a Conferenza Internazionale sull'Apprendimento Automatico (pp. 1691-1703). PMLR. available at: <https://proceedings.mlr.press/v119/chen20s.html>
- Eco, U. 1984 . *Semiotica e filosofia del linguaggio*, Einaudi, Torino.
- Fernández Gambín, A., Yazidi, A., Vasilakos, A. et al., 2024. Deepfakes: current and future trends. *Artificial Intelligence Review*, 57:64.
- Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., Ureña-López, L. A. 2021 . A survey on bias in deep NLP. *Applied Sciences*, 11(7), 3184.
- Groupe  $\mu$  (1970). *Rhétorique générale*, Larousse, Paris (trad. it. *Retorica generale*, Bopiani, Milano 1976; II ediz. 1980)
- Groupe  $\mu$  1992 . *Traité du signe visuel*. Seuil, Paris.
- Ho, J., Jain, A., Abbeel, P. 2020 . Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- Kaur, D., Kaur, Y. 2014 . Various image segmentation techniques: a review. *International Journal of Computer Science and Mobile Computing*, 3(5), 809-814.
- Lakoff, G., Johnson, M. 1980 . *Metaphors we live by*, The University of Chicago Press, Chicago (trad. it. *Metafora e vita quotidiana*, Bompiani, Milano 1998).
- LeCun, Y., Bengio, Y., Hinton, G. 2015 . Deep learning. *nature*, 521(7553), 436-444.
- Prystawski, B., Thibodeau, P., Potts, C. et al., 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.
- Ramesh, A., Dhariwal, P., Nichol, A., et al., 2022. Generazione di immagini gerarchica condizionata al testo con latenti CLIP. ArXiv, cs.CV. available at: <https://arxiv.org/abs/2204.06125>
- Rombach, R., Blattmann, A., Lorenz, D. et al., 2022 . Sintesi di immagini ad alta risoluzione con modelli di diffusione latenti. ArXiv, cs.CV. available at: <https://arxiv.org/abs/2112.10752>
- Sonesson, G. 1989 . *Pictorial concepts*, Lund University Press, Lund.
- Tong, X., Choenni, R., Lewis, M. et al., 2024. Dataset di sfida per la comprensione delle metafore per i modelli di linguaggio di grandi dimensioni. ArXiv, cs.CL. available at: <https://arxiv.org/abs/2403.11810>
- Yakura, H. 2023 . Evaluating large language models' ability to understand metaphor and sarcasm using a screening test for Asperger syndrome. *arXiv preprint arXiv:2309.10744*.
- Zoph, B., Raffel, C., Schuurmans, D. et al., 2022 . Abilità emergenti dei grandi modelli di linguaggio. TMLR.