UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZE POLITICHE E DELLA COMUNICAZIONE

CORSO DI LAUREA MAGISTRALE

IN

CORPORATE COMMUNICATION E MEDIA

TESI DI LAUREA

IN

GESTIONE DELLA CONOSCENZA PER I SERVIZI DIGITALI

Evaluating the effectiveness of neural language models on commonsense categorization: a comparison with the Dual-PECCS system

Relatore:

Ch.mo Prof.
Antonio Lieto

Candidata:

Isabella Cossidente
Matr.: 0323101547

ANNO ACCADEMICO 2023/2024

Abstract

This research work evaluates the accuracy of different neural language models in conceptual categorization tasks by comparing the results of different systems with a gold standard obtained by the responses of human subjects. The models considered are Open AI's ChatGPT-3.5, ChatGPT-4 and 4o, and Microsoft Copilot, which were compared with Dual-PECCS, a cognitively inspired knowledge representation and reasoning system developed by Antonio Lieto, Daniele P. Radicioni and Valentina Rho. The investigation was also extended to Open AI's DALL·E 3, to evaluate if and to what extent the created images fit the expected conceptual representation.

Each model was presented with linguistic descriptions used in Dual-PECCS, referring to prototypes or exemplars of animals. The performances were evaluated following two metrics, likewise taken from Dual-PECCS: concept-categorization accuracy and proxyfication accuracy, indicating the effectiveness in retrieving the expected concept and the expected proxyfied representation, respectively.

As for text generation, the calculated rates showed that GPT models outperform Microsoft Copilot, but all systems do worse than Dual-PECCS in general. For what concerns images, just over half of the results showed a match in conceptual categorization, but still not free from common fails committed by Artificial Intelligence image generators. In most cases, the main errors detected in the systems' outputs regard the retrieving of a wrong exemplar, in terms of both text and images.

Through the research, we can gain insight into the potential and limitations of current Large Language Models.

Keywords: conceptual categorization, Dual-PECCS, language models, prototypes, exemplars

Acknowledgments

I would like to acknowledge the many people whose presence has really made a difference through this journey.

I want to express my thanks to my advisor, Professor Antonio Lieto, for his guidance and encouragement throughout this past year. His mentorship has not only been important for my thesis, he has also taught me how to embrace change, still recognizing the enduring nature of our human consciousness, amid technological progress.

I'm also deeply grateful to my family for their support and for understanding of my choices. Their acceptance has given me the confidence to pursue my path without feeling judged, which is a privilege not all daughters, sons or students experience.

Finally, a special thanks to my closest friends, both those who have been with me from the beginning and those who started as colleagues and have become so much more. Thank you for supporting me during the highs and the lows.

Table of Contents

List of Figures

List of Tables

Introduction

The cognitive process of categorization is something that human beings naturally engage in to mentally organize the elements of the world. Our recognition experience is simplified by grouping entities based on common characteristics: we can identify and keep recurring traits, extend them to new entities and draw inferences about their properties. The studies made in Cognitive Science have provided multiple theories on how humans organize and retrieve conceptual information. There are two primary positions. The initial theory, which is referred to as classical or Aristotelian, states that the meaning of concepts can be determined according to a set of necessary and sufficient conditions. The classical approach has been the basis of other theories that lasted until the 1980s and fall under the so-called componential analysis. This position holds that the meaning of concepts is determined by the sum of a variety of binary [1] semantic features, required for the members of a given category. [2]

The counter-theory of prototypical analysis, that began with the experimental results of the American psychologist Eleanor Rosch in the mid 1970s, has demonstrated the inadequacy of the componential approach. Rosch's prototype theory suggests that concepts are arranged in our minds as prototypes, that is, the best instances of a category.

Another theoretical construct is the exemplar theory. It states that we mentally perceive a certain category as a collection of exemplars we have encountered throughout our experience, and that we have stored in our memory; the exemplars represent the benchmark for the categorization of new elements [2].

A series of follow-up studies, starting from Malt [3], has shown that there is no mutual exclusion between prototypes-based and exemplars-based classification, since subjects do not employ one single categorization strategy. Depending on a number of factors, such as the nature of the

[1] The binary value of a semantic feature means it is either present (+) or absent (-) in the definition of a concept. Example of componential analysis of the concept *Woman*: ANIMATE, HUMAN, FEMALE, ADULT.

[2] Using the concept *fruit* as en example, the prototypical representation we would have in mind would coincide with an apple, which possesses many of the characteristics we commonly associate with fruit, such as sweetness, juiciness, freshness, has peel and seeds. If we employ the exemplar theory, when we come across an unknown fruit, we will compare it to all the fruit exemplars we already know to decide whether it belongs to the same category.

stimuli, the type of task or the type of reasoning — non-monotonic versus standard deductive reasoning — one strategy may be dominant, but they can also coexist and be used alongside classical representations. Concepts are therefore conceivable in terms of heterogeneous groups of information and mental representations [4, 5].

Speaking of Artificial Intelligence, perceptual tasks, including categorization itself, are well performed by neural networks. The Large Language Models (LLMs) discussed in this work belong to this group of neural AI systems. LLMs are designed to understand and generate human language. They are trained on vast amounts of textual data, enabling them to perform tasks such as text generation, translation, and summarization. These models are pivotal in applications like virtual assistants and automated content creation. Simple Recurrent Networks (SRNs), introduced by Jeff Elman in 1990 [6], are the earliest neural network architectures, designed to handle sequential data by maintaining a form of memory over time. Systems like GPT and DALL·E owe their development to SRNs. These are transformers [7], highly complex neural networks that are trained to predict sequences of words that follow input data, using probability calculations based on their training set. Also, transformers have self-attention mechanisms that enable them to process input data globally and simultaneously, instead of sequentially.

In the evolving field of generative AI, understanding the capabilities and limitations of these neural models is crucial. The computational power of these systems enables them to achieve a general accuracy but, because they lack of real understanding, meanings and contexts may be misinterpreted. This can lead to errors, especially in tasks requiring common sense reasoning. In fact, these kinds of models are designed according to a functionalist approach: they are able to replicate cognitive processes using completely different mechanisms, which are often inexplicable. They are far from being biologically relatable, as structuralist models may be [5].

Based on these premises, this work aims to evaluate the accuracy of conceptual categorization tasks performed by these systems, considering both text-to-text and text-to-image. Furthermore, the study looks at a comparison with the Dual-PECCS computational categorization model, which is based on assumptions of all theories mentioned above: prototypical reasoning,

exemplars-based reasoning, and standard monotonic categorization procedures. The connecting link between these various reasoning approaches integrated in Dual-PECCS is the dual process theory of reasoning, postulated by Daniel Kahneman [8].

The thesis is structured into three main chapters. The first chapter presents the Dual-PECCS system and its outcomes in the categorization process, which set up the basis for the development of this project. The second chapter provides a detailed explanation of the methodology adopted for the research to measure the performance of categorization tasks in different neural language models, including the research design, data collection methods, and analytical techniques used in the study. The third chapter presents the findings of the research and the subsequent comparison of the systems with Dual-PECCS to discuss their effectiveness.

Chapter 1

An overview of Dual-PECCS

1.1 The Dual Process Theory of Thinking

Before moving forward with the discussion of Dual-PECCS, it is important to mention another of its foundational elements, connected to the categorization procedures of the system: the theory of the dual process advanced by the Israeli psychologist Daniel Kahneman, winner of the Nobel Prize for Economics in 2002. In his work *Thinking, Fast and Slow* [9] Kahneman talks about two systems of the mind, called System 1 and System 2, that oversee two types of processes, one automatic and one effortful.

System 1 handles all those activities and innate skills that require no effort and no voluntary control. It operates automatically and it is always active and running. Activities associated with System 1 are, for example, reading or driving on an empty road, but it also linked to associations between ideas that we have learnt and then come to our mind involuntarily. An example given by Kahneman is that we cannot help but think of Paris when the capital of France is mentioned.

System 2 governs mental activities that require attention, effort and concentration, therefore they are too complex to be conducted at once. While System 1 is automatic, System 2 is slow and controlled. Examples of this kind of activities are focusing on someone's voice in a crowded room or counting the occurrences of a particular letter in a written text.

The two systems interact with each other, optimizing people's performances, since they are at the basis of many cognitive processes: System 1 creates unconscious impressions and feelings, System 2 is associated with deliberate choices; moreover, the control wielded by System 2 is useful to hold back unruled impulses of System 1. The following presentation of Dual-PECCS will offer an understanding of how this theory is integrated into it.

1.2 Dual-PECCS

Dual-PECCS (**P**rototypes and **E**xemplars **C**onceptual **C**ategorization **S**ystem) is a cognitively inspired categorization system[1], first presented in 2015 [10], developed through the collaboration of Antonio Lieto, Daniele P. Radicioni and Valentina Rho at the University of Turin, in Italy. It is now possible to explain how the assumptions mentioned earlier are integrated into the system's conceptual architecture. There is an explicitly heterogeneous modeling that encompasses prototypical and exemplary representations, as well as classical ones. WordNet is the linguistic resource that connects all these pieces of information about entities[2].

The dual process theory is used to coordinate different reasoning strategies. In Dual-PECCS, prototypes and exemplars-based categorization, linked to commonsense representation, is associated with the typical fast processes of System 1, while classical representations, referring to standard deductive logic, are associated with the processes of System 2, which are slower.

The architecture is a hybrid of two frameworks: conceptual spaces and ontologies. Conceptual spaces, proposed by Peter Gärdenfors in 2000, depict commonsense knowledge as a geometric structure. In these spaces, concepts are qualified as a set of quality dimensions that can refer either to perceptual or abstract information. Semantic similarities between entities, and the resulting typicality effects, are extracted through a series of calculations. When given a concept that corresponds to a precise geometrical region in the space, each point that falls within that region has a certain degree of centrality. Prototypes are related to the geometric center of a convex region (referred to as a centroid), while exemplars are related to the intermediate distance between the two points [12]. The other component of automatic reasoning is managed through ontologies, a form of knowledge representation, specifically symbolic. Dual-PECCS integrates

[1] Cognitively inspired systems take inspiration from psychological theories to replicate cognitive mechanisms such as attention, reasoning or memory. They differ from biologically inspired systems, which are based on Neuroscience studies and aim to replicate the physical and functional structure of the human brain, using artificial neural networks.

[2] WordNet is a lexical database for the English language, developed at Princeton University. Words' organization is based on semantic relations through groups of synonyms called *synsets*, which represent distinct concepts. Synsets are also linked through taxonomic classifications, with inclusion and subordination relations [11].

Cyc (1984) by Douglas Lenat, one of the most exhaustive knowledge bases, to treat classical representations. Figure 1 below displays how the heterogeneous representation of a concept is structured: it encloses the prototypical representation of *tiger*, the exemplary representation of *white tiger*, and the classical representation of the concept.

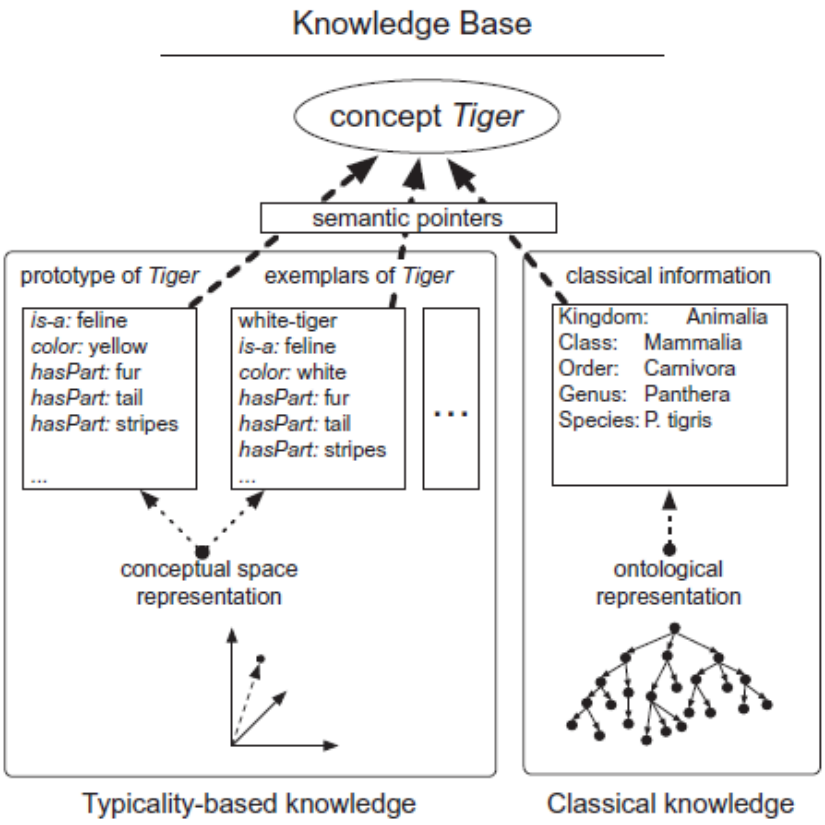

Figure 1: Representation of the concept *Tiger* in Dual-PECCS [1]

Dual-PECCS has been added as an external system to the cognitive architectures ACT-R (Anderson et al., 2004) and SOAR (Laird, 2012), expanding their ability to represent and process knowledge. A different suggestion was made to connect representations of prototypes and exemplars with the assumptions of the so-called theory-theory, always maintaining an heterogeneous perspective. According to the theory-theory approach, the comprehension of things is based on their relationship to each other, resulting in complex mental concepts. Items are

categorized by using broader concepts that incorporate general knowledge about a given category. The proposed solution consists of a new classification algorithm known as DELTA (unifie**D** cat**E**gorization a**L**gorithm for he**T**erogeneous represent**A**tions). DELTA better harmonizes all these different ways of categorizing information, for a more realistic representation of knowledge in artificial cognitive agents. The algorithm categorizes stimuli by first checking for similarity with stored exemplars — it prioritizes exemplars over prototypes to mirror human categorization behavior —; if none are found, it looks for the closest prototype. Additionally, it can use theory-like structures for a more informed decision-making, especially when a stimulus presents conflicting traits [13].

1.3 Test and Evaluation

Dual-PECCS was tested on concept retrieval and commonsense reasoning on the basis of a dataset composed of 112 linguistic descriptions, which all referred to concepts of animals. The dataset was specifically developed by a team of neuropsychologists, linguists and philosophers. Each description is similar to a short riddle and is related to an expected target answer, either a prototype or an exemplar (e.g. "The rodent that eats cheese", expected result: *mouse*, expected type: *prototype*; or "The blue amphibian that lives in a lake", expected result: *blue frog*, expected type: *exemplar*). The expected results are a so-called gold standard, since they correspond to answers given by human subjects in an experiment conducted at the University of Turin, where participants were asked to perform a task of naming from definition, under different conditions. Dual-PECCS' test involved two experimental settings, one of which included Information Extraction (IE) where the entire process was conducted from input to output, beginning with the textual description as initial input. An alternative approach, which does not rely on Information Extraction (no IE), is to manually convert the input into a chunk request, based on valuable information for each concept.

Two metrics were used to evaluate the performance of the system: concept-categorization accuracy (CC-Acc) and proxyfication accuracy (P-Acc). Given an input, concept-categorization

accuracy measures how accurately the system can identify and categorize the intended concept. The retrieval is considered accurate even if the target representation is incorrect. Proxyfication accuracy measures how well the system can retrieve the specific proxyfied representation. In such case, retrieving the category is not enough: if the system confuses different types of representations, it is counted as an error (fig. 2, table a). Proxyfication errors can be of three kinds (fig. 2, table b): Ex-Proto, if an exemplar is returned instead of the expected prototype; Proto-Ex, if a prototype is returned in place of an expected exemplar; Ex-Ex, if the system retrieves a mistaken exemplar [1].

a. Accuracy rates obtained for the conceptual categorization accuracy (CC-ACC) and proxyfication accuracy (P-ACC) metrics.

Test	CC-ACC	P-ACC
With no IE	89.3% (100/112)	79.0% (79/100)
With IE	77.7% (87/112)	71.3% (62/87)

b. Analysis of the errors in the proxyfication (P-ACC metrics).

	Proxyfication error		
Test	Ex-Proto	Proto-Ex	Ex-Ex
With no IE	21.0% (21/100)	0.0% (0/100)	0.0% (0/100)
With IE	28.8% (26/87)	0.0% (0/87)	5.8% (5/87)

Figure 2:
Dual-PECCS' results extracted from [1]

The table above illustrates the effectiveness of Dual-PECCS both with and without Information Extraction. An error analysis is provided too. In 3.3, we will discuss these results in greater detail and compare them to the various neural systems. However, it may be noted that these percentages in commonsense reasoning are quite high and satisfactory. They not only compare favorably to the gold standard (i.e., human performance) but also outperform other systems that lack understanding, such as Google Translate or Watson [14].

1.4 Minimal Cognitive Grid

The Minimal Cognitive Grid is an evaluation model based on three principles useful for placing a system on a more structuralist or functionalist design axis, and it is applicable to both

biologically and cognitively inspired systems. Antonio Lieto, its inventor, defines it as "a non subjective, graded, evaluation framework allowing both quantitative and qualitative analysis about the cognitive adequacy and the human-like performances of artificial systems in both single and multi-tasking settings" (2021). The three dimensions included in the Grid are:

1. **Functional/Structural Ratio**: the complete system is analyzed through a dissection to determine how many elements have functionalist computation modes and how many have structuralist modes, just by counting them;

2. **Generality**: the objective of this dimension is to evaluate the extent to which a system can be used in various tasks, whether it can replicate multiple cognitive functions or only some. Here, as well, we can count how many cognitive faculties can be modeled within a single system;

3. **Performance Match**: there is the comparison between the outcomes of natural systems and artificial systems, not only in terms of results, but also including errors and execution time, which should be close to those obtained by humans. This increases the precision of calculating the system's level of plausibility, even though a good performance match does not necessarily mean the model is structuralist.

Along with the discussion of outcomes of the systems, the Grid will be used in 3.3 to carry out a multidimensional comparison between the various models examined.

Chapter 2

Research Methodology

2.1 Research Design

The categorization results of Dual-PECCS provide the foundation for this work, which in fact serves as a comparative research to observe how well newest Large Language Models can perform in categorizing prototypes and exemplars, and if their accuracy rates match with those calculated in Dual-PECCS, or if they supposedly do better.

The evaluation of conceptual categorization was extended to both text and image generation, to see how linguistic descriptions are creatively interpreted for concept retrieving. For text-to-text testing, the chosen models are Open AI's ChatGPT-3.5, ChatGPT-4 and 4o, and Microsoft Copilot, while for text-to-image generation DALL·E 3 technology was used, integrated in Microsoft Designer.

2.2 Data Collection

Data were collected by submitting each language model to the 112 experimental descriptions previously used in Dual-PECCS. A brief explanation of the use of Microsoft Copilot is required. The system offers three distinct conversation styles: a more creative one, that generates original answers; a more balanced one, a default style that provides useful and interesting information; a more precise one, that focuses on rigorous and detailed knowledge. For the research, the model was set on the more balanced conversation style, which I thought would be the most compatible with human commonsense reasoning strategies, being flexible enough to achieve a good ratio of precision and engagement. Beyond the style that was selected, all systems have presented cases where the answer was more straightforward, and others with more potential alternatives,

even actual lists of numbered concepts that could possibly match the given input. Among these alternatives there could be the correct answer, or they could all be wrong. An example is shown in the following figure 3.

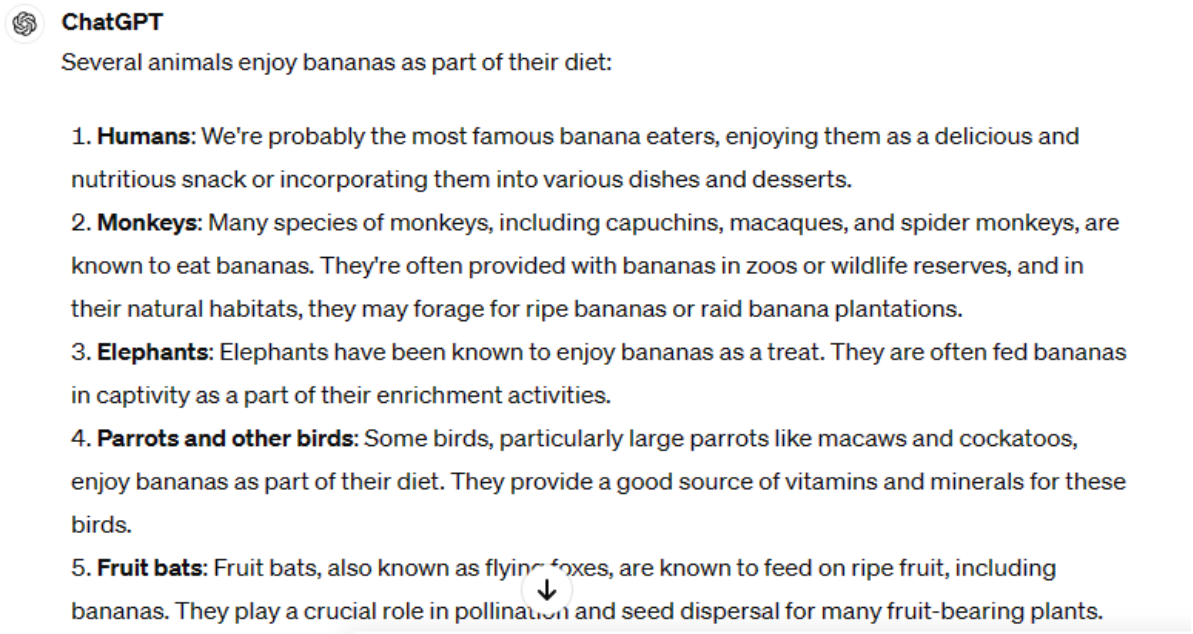

Figure 3: ChatGPT-3.5 answering the input text *The animal that eats bananas* (Expected result: Monkey)

The same descriptions were used as input text for the creation of images in DALL·E 3. The expression was given to the model as it was, so it did not receive any suggestion about the possible style of the representation. According to this, the system generated more realistic pictures (fig. 4), while others were created following a cartoonized or caricatural style (fig. 5) For each prompt, the model generated 4 different images, for a total of 448 pictures. The images that depict the same concept are typically consistent with each other, adapting to the same style, or at the very least to a similar one.

Figure 4: DALL·E 3 generated the image for the input text *The very slow animal with one feeler and a shell*. (Expected result: Snail Exemplar One Feeler)

Figure 5: DALL·E 3 generated the image for the input text *A black bird with yellow beak* (Expected result: Blackbird)

2.3 Data Analysis

All data were put in a table and analysed according to the two metrics originally designed for Dual-PECCS, that is, concept-categorization accuracy and proxyfication accuracy. The results given by the models were evaluated assigning a binary value between 0 and 1 for each metric, depending on the correspondence of the concept. There are three possible scenarios to consider:

- if both CC-Acc and P-Acc are equivalent to 0, the concept has to be considered as a total error, since neither the concept nor the representation was retrieved (table 1);
- if the retrieved concept is correct but the expected representation is not, CC-Acc is equivalent to 1, while P-Acc is still equivalent to 0 (table 2). This scenario gives rise to the three further types of proxyfication errors, already presented in 1.3;
- if both the retrieved concept and the retrieved representation match with the expected outcome of the description, both values are equivalent to 1 (table 3).

Input text	Expected result	Expected type	ChatGPT-3.5	CC-Acc	P-Acc
The bird with black fur with finned paws and that is able to swim	Black Swan	Exemplar	Penguin	0	0

Table 1: Concept evaluation in ChatGPT-3.5

Input text	Expected result	Expected type	Copilot	CC-Acc	P-Acc
The black and brown animal with eight striped legs	Tarantula spider	Exemplar	Brown widow spider	1	0

Table 2: Concept evaluation in Microsoft Copilot

Input text	Expected result	Expected type	ChatGPT-4	CC-Acc	P-Acc
The insect with sting and black and yellow striped body that produces honey	Bee	Prototype	Honeybee	1	1

Table 3: Concept evaluation in ChatGPT-4

After classifying all the concepts, those of the second type were checked out again to examine proxyfication errors, distinguishing between Ex-Proto, Proto-Ex, and Ex-Ex, according to the representation given by the model.

Input text	Expected result	Expected type	ChatGPT-4	Proxyfication error
The raptor with big wings that flies over mountains	Eagle	Prototype	Golden eagle	Ex-Proto
The fish that lives in freshwater lakes or rivers found naturally in areas close to the Adriatic Sea	Adriatic Trout	Exemplar	Trout	Proto-Ex
The equine with long red ears and with white and black fur	Catalan donkey	Exemplar	Pinto donkey	Ex-Ex

Table 4: Proxyfication errors in Microsoft Copilot

The same process was employed to categorize the images one by one, since not all four representations of a given concept share the same values.

The sum of all these data has led to a series of results on the performance of the cited models, discussed in chapter 3, as well as the comparison with the Dual-PECCS system.

Chapter 3

Results

3.1 Text-to-Text Performance

As established by the metrics, concept-categorization accuracy is calculated on the total number of concepts (which is 112), while proxyfication accuracy and its errors, as subsets of the former, are calculated on the basis of the number of concepts reclaimed (which is, in fact, concept-categorization accuracy itself).

The system that shows the highest accuracy rates is ChatGPT-4o with an accuracy of 82.1% in concept-categorization and 81.5% in proxyfication. In decreasing order, ChatGPT-4 follows, with not much lower rates, respectively of 80.3% and 75.5%. Then, there is ChatGPT-3.5, with a concept-categorization accuracy of 70.5% and a proxyfication accuracy of 76.0%. Lastly, Microsoft Copilot shows a concept-categorization accuracy rate of 65.0% and a proxyfication accuracy rate of 74.0%.

Despite ChatGPT-3.5 having a lower overall conceptual retrieval capacity than GPT-4, it was able to retrieve accurate representations more effectively than its more advanced model, but not more than GPT-4o. GPT-4o (which stands for 'omni') is a significant improvement over previous versions, particularly in visual and audio comprehension. When it comes to text and reasoning, GPT-4o reaches the same performance level as GPT-4 Turbo (a version that is even more optimized than 4). In any case, it is about twice as fast and it is cheaper to program its features [15]. The difference is minimal when compared to concept-categorization accuracy, but the disparity in proxyfication accuracy is more noticeable. In the proxyfication task, Microsoft Copilot displays rates that are relatively comparable to those of other models, although conceptual categorization is much lower.

When it comes to proxyfication errors, the most common type is Ex-Ex in GPT models – 12.7%

in GPT-3.5, 13.3% in GPT-4 – and Proto-Ex in Copilot and GPT-4o – with 12.3% and 10.9%. The data are sorted in more detail in the tables that follow.

Model	**CC-Acc**	**P-Acc**
ChatGPT-3.5	70.5% (79/112)	76.0% (60/79)
ChatGPT-4	80.3% (90/112)	75.5% (68/90)
ChatGPT-4o	82.1% (92/112)	81.5% (75/92)
Microsoft Copilot	65.0% (73/112)	74.0% (54/73)

Table 5: Accuracy rates in texts

Model	**Ex-Proto**	**Proto-Ex**	**Ex-Ex**
ChatGPT-3.5	6.3% (5/79)	6.3% (5/79)	12.7% (10/79)
ChatGPT-4	4.4% (4/90)	6.7% (6/90)	13.3% (12/90)
ChatGPT-4o	1% (1/92)	10.9% (10/92)	7.7% (7/92)
Microsoft Copilot	6.8% (5/73)	12.3% (9/73)	6.8% (5/73)

Table 6: Proxyfication errors in texts

3.2 Text-to-Image Performance

According to the data on image creation, concept-categorization accuracy has a value of 53.3%, covering just over half of the 448 images. Nevertheless, the value of proxyfication is quite high, with an accuracy of 85.8%. As with text generation models, the most common proxyfication error in this case concerns the exchange between exemplars.

CC-Acc	**P-Acc**
53.3% (239/448)	85.8% (205/239)

Table 7: Accuracy rates in DALL·E 3

Ex-Proto	Proto-Ex	Ex-Ex
2.1% (5/239)	5.0% (12/239)	8.7% (21/239)

Table 8: Proxyfication errors in DALL·E 3

The evaluation of some results may be challenging, due to ambivalence in several pictures. Many of these images are a product of hallucinations, which are commonly experienced by Artificial Intelligence systems. The term "hallucination" refers to cases where a model delivers a false or incorrect answer, but it is presented as plausible and convincing within the context, ultimately being misleading [16]. In text-to-image models, hallucinations occur when they generate images that do not match the textual description given, which is misinterpreted because of ambiguities, insufficient training and machine limitations [1]. Recurring examples of hallucinations in visual representations involve adding elements that were not included in the input text, or deleting others, distorting shapes, colors, and proportions.

In this work, hallucinations were found mainly in representations classified as total errors, where neither the concept nor the representation were retrieved. Among these errors lie two possibilities: some descriptions have been interpreted by simply portraying a different but existing animal, while others have given rise to images of completely invented beings or animals. It is advantageous for the model that the former kind or representations are much greater in number than the latter kind (counting 141 vs 68). Examples are given below.

The representation in figure 6 was classified as an error since it does not match with the expected result *Hippo*, as specified in the caption. However, the subject of the picture is easily recognizable: it clearly depicts an elephant. The hallucination effect can be seen in the fact that the animal has three tusks, instead of two.

[1] As previously stated, generative AI models' functioning is based on words prediction mechanisms, so they just generate plausible content, which is not necessarily true. Besides, limits often are at the root of the process, that is, if training data is inaccurate - as the Internet is filled with it - the models will reproduce inaccuracies [17].

Figure 6: DALL·E 3 generated the image for the input text *The big mammal that is herbivore and lives in the savanna and swims*. (Expected result: Hippo)

(a)

(b)

Figure 7: DALL·E 3 generated the images for the input text *A large mammal with long claws that hunts fish in mountain rivers*. (Expected result: Bear Fish Hunter)

Figure 7, gathers two images created from the same input text. These are evidently unreal animals, resulting from the model hallucinating. Considering the expected result *Bear Fish Hunter*, representation (a) is far from any resemblance, since it depicts some sort of moose with non-existent tusks; representation (b) actually portrays a bear, although it has unnaturally long claws and, once again, two fictitious tusks.

The same lack of semantic understanding and contextualization that was mentioned for text-to-text generation is also present in text-to-image generation. Another example is provided by figure 8, whose input description was *The big mammal with white fur that lives in Arctic and that eats walruses*. The image was completely misrepresented by the walrus being interpreted as a subject instead of being the object. Like other AI models, DALL·E is trained on vast datasets containing images and descriptions, however, they may not always perfectly understand specific nuances. Sometimes it might not have enough context to generate what was envisioned.

Figure 8: DALL·E 3 generated the image for the input text *The big mammal with white fur that lives in Arctic and that eats walruses*. (Expected result: Polar Bear)

3.3 Comparison with Dual-PECCS

The data about accuracy results from each model's performance are summarized in table 9 and table 10 below, which include Dual-PECCS as well and its tests made with and without Information Extraction.

Model	**CC-Acc**	**P-Acc**
Dual-PECCS (no IE)	89.3%	79.0%
ChatGPT-4o	82.1%	81.5%
ChatGPT-4	80.3%	75.5%
Dual-PECCS (with IE)	77.7%	71.3%
ChatGPT-3.5	70.5%	76.0%
Microsoft Copilot	65.0%	74.0%
DALL·E 3	53.3%	85.8%

Table 9: Overall accuracy rates

It is noteworthy that Dual-PECCS stands out in the highest positions. Particularly, its component without IE turns out to be the best system in terms of concept-categorization accuracy, while the component with IE is overcome only by GPT-4 variants. All the models left have a worse performance than Dual-PECCS in both cases. The data for proxyfication accuracy are more consistent and closer to each other. GPT-4o's ability to recover exact representations with great precision outperforms Dual-PECCS in this area. Also GPT-3.5 is above GPT-4, while Microsoft Copilot qualifies as the last among text-to-text models, for both metrics, despite its integration in several tools of Microsoft 365 and, recently, also in WhatsApp [18]. As for DALL·E 3, it has the highest percentage among all models in proxyfication accuracy.

Finding such "poor" data for these deep learning models is quite surprising. Despite the high costs required for their training — both economic and enviromental — and updating, they are still inferior to a system that is a bit more outdated and computationally less expensive. GPT-4 and 4o models are the most obvious examples of the trade-off between cost and efficiency. The

significant investment made in these systems (since training GPT-4 cost more than $100 million [19]) ensures faster execution and effective management of input information. Specifically, ChatGPT-4o is multimodal and can answer to audio inputs in an average time of 320 milliseconds, which is similar to human response time in a conversation [15]. In general, as stated in 3.1, it is designed to mitigate impacts by improving effectiveness and reducing resource consumption.

In reference to 1.4, a comparative analysis of of the systems can be made according to the parameter of performance match of the Minimal Cognitive Grid. This parameter is the the most relevant one for the kind of research conducted, however we can consider also the other two. As for the functional/structural ratio, DUAL PECCS' ratio is lower, therefore the system is more accurate from a structural point of view, being able to combine different forms of reasoning. The other systems involved heavily lean towards functionalist computation. They process inputs through neural networks designed for language and image modeling. There is not any structuralist framework that dictates behavior beyond the learned patterns.

For what concerns the criterion of generality, Dual-PECCS surely has to be considered a general system when it comes to categorization tasks. The examined systems – except for DALL·E 3, whose ability is restricted to image generation – exhibit more task versatility. They can replicate multiple cognitive functions and engage in various tasks, such as answering questions, generating creative content in form of text, images and voice, summarizing information, engaging in conversation, all within a single framework.

Moving forward with the performance match, comparing the effectiveness of Artificial Intelligence with that of humans is interesting. It has already been outlined that Dual-PECCS has excellent proportions compared to the responses given by human subjects. Nonetheless, the performance match also considers the time and errors that the system makes, which, in general, may or may not be as close to the level of human performance. The efficiency of the Conceptual Spaces, characterized by outstanding complexity, results in the execution time of Dual-PECCS being quite fast, often less than a second [5]. The other systems also have a quite fast prompt

response times, within a few seconds or less. As stated above, GPT-4o has a very good yield compared to human performance. DALL·E 3 can take a bit longer due to image generation, usually around 10 seconds, depending on the complexity of the image requested.

Model	**Ex-Proto**	**Proto-Ex**	**Ex-Ex**
Dual-PECCS (with IE)	28.8%	0.0%	5.8%
Dual-PECCS (no IE)	21.0%	0.0%	0.0%
ChatGPT-3.5	6.3%	6.3%	12.7%
ChatGPT-4	4.4%	6.7%	13.3%
ChatGPT-4o	1%	10.9%	7.7%
Microsoft Copilot	6.8%	12.3%	6.8%
DALL·E 3	2.1%	5.0%	8.7%

Table 10: Overall proxyfication errors

With respect to proxyfication errors, one particular condition is that Dual-PECCS has a higher concentration of errors within a specific typology (Ex-Proto), whereas all other systems have a more even distribution of errors across the three typologies. Most errors made by Dual-PECCS are caused by the system's tendency to confuse exemplars and prototypes, returning the former instead of the latter. The system tends to favor exemplars that have detailed information that match the description, even though it is counterintuitive when a very general description is given. Ex-Ex errors have a lower error rate, mainly due to confusing characteristics extracted from linguistic descriptions [1]. ChatGPT-4o, which has been noted to have the highest rank among all other systems, has oddly opposite percentages to those of Dual-PECCS: given the same input descriptions, the model is more susceptible to mistakes by retrieving the prototype instead of the exemplar (Proto-Ex), while the percentage of of Ex-Proto is minimal. Proto-Ex errors are also predominant in Microsoft Copilot but, generally speaking, the most common type of error is the retrieving of a different exemplar than the expected one (Ex-Ex). This may be due to a few key reasons: Large Language Models generate human-like language, but they

rely on statistical patterns they have been trained on, rather than genuine understanding. They learn from vast amounts of text, so if some exemplars are frequently presented in similar contexts, the model might make incorrect associations.

Regarding the total errors (those with values 0 and 0, table 1) that lower the CC-Acc rate, it should be said that in text-to-text models there were no cases of nonsensical replies, or replies completely far from the input description. Even though these are still inaccuracies, the hallucinations discussed earlier have been found primarily in the text-to-image model.

Ultimately, these models show a decent performance match, but discrepancies in accuracy and error rates indicate that it does not properly replicate human commonsense categorization and, despite the massive work behind Large Language Models, their overall execution is worse than Dual-PECCS' [2].

[2] All the data analyzed and presented in this thesis are available at this URL: https://drive.google.com/file/d/1OOOQYsS99eJ2Uv5r7ie0boTsboJDjGe0/view?usp=drive_link.

Conclusion

The present work is focused on a research that reveals the varying degrees of accuracy and the limitations that can be found in different neural language models, when tasked with conceptual categorization. The study was centered around Open AI's ChatGPT-3.5, ChatGPT-4 and 4o, Microsoft Copilot for text-to-text generation, and Open AI's DALL·E 3 for text-to-image generation. The work not only enlightens the potential of these models, but also emphasizes their limitations in contrast with the cognitively-inspired system Dual-PECCS. The system as presented and outlined in its key points for the discussion in Chapter 1.

Chapter 2 provides information on the methodology used to collect data from the chosen neural models, as well as the parameters required for their analysis and understanding.

Chapter 3 contains a dissertation on research findings. In spite of the advancements in the field of Large Language Models, and generative Artificial Intelligence in general, though partially successful, these models still struggle to replicate human commonsense categorization effectively and they are not able to reach the categorization accuracy rates achieved by Dual-PECCS.

The Dual-PECCS system has remained open for further research, but basically it serves as a practical model for how these ideas can be applied computationally. Certainly, the considered models are promising in the functions they have been designed for; however, significant gaps remain in their ability to replicate the nuances of human commonsense reasoning.

Ultimately, our findings suggest a need for ongoing refinement and development within the field: future investigations should be aimed at enhancing AI's understanding of concepts and improving its effectiveness in tasks requiring commonsense reasoning.

References

[1] A. Lieto, D. P. Radicioni, and V. Rho, "Dual PECCS: a cognitive system for conceptual representation and categorization," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 29, no. 2, 2017.

[2] C. Cacciari, *Psicologia del linguaggio, Second Edition*. il Mulino, 2011. pp. 153-155.

[3] B. C. Malt, "An On-Line Investigation of Prototype and Exemplar Strategies in Classification," *Journal of Experimental Psychology Learning Memory and Cognition*, 1989.

[4] A. Lieto, "A Computational Framework for Concept Representation in Cognitive Systems and Architectures: Concepts as Heterogeneous Proxytypes," *Procedia Computer Science*, vol. 41, 2014.

[5] A. Lieto, *Cognitive Design for Artificial Minds*. Routledge, 2021. pp. 20-24, 66-69, 75-76.

[6] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, 1990. pp. 179-211.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv*, 2017.

[8] A. Lieto, D. P. Radicioni, and V. Rho, "Dual-PECCS." www.dualpeccs.di.unito.it Accessed: 2024-09-25.

[9] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux Inc, 2013. pp. 22-27.

[10] A. Lieto, D. P. Radicioni, and V. Rho, "A Common-Sense Conceptual Categorization System Integrating Heterogeneous Proxytypes and the Dual Process of Reasoning," *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[11] Princeton University, "WordNet. A Lexical Database for English." wordnet.princeton.edu. Accessed: 2024-09-29.

[12] A. Lieto, A. Chella, and M. Frixione, "Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation," *Biologically inspired cognitive architectures*, vol. 19, pp. 1–9, 2017.

[13] A. Lieto, "Heterogeneous proxytypes extended: Integrating theory-like representations and mechanisms with prototypes and exemplars," *Biologically Inspired Cognitive Architectures 2018: Proceedings of the Ninth Annual Meeting of the BICA Society*, pp. 217–227, 2019.

[14] E. Davis and G. Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM*, vol. 58, issue 9, 2015. pp. 92-103.

[15] Open AI, "Hello GPT-4o." https://openai.com/index/hello-gpt-4o/.
Accessed: 2024-09-29.

[16] Merriam-Webster, "Hallucination." www.merriam-webster.com/dictionary/hallucination
Accessed: 2024-09-26.

[17] MIT Sloan Teaching and Learning Technologies, "When AI Gets It Wrong: Addressing AI Hallucinations and Bias." www.mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias. Accessed: 2024-09-26.

[18] Microsoft, "Copilot for social apps." https://support.microsoft.com/en-gb/topic/copilot-for-social-apps-43eb625d-eb25-4c72-a458-19842bf42212. Accessed: 2024-10-08.

[19] W. Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," 2023. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.