Università
Ca’Foscari
Venezia

Master’s Degree programme in

Digital and Public Humanities

Final thesis

Inquiring the ‘oracle’

An empirical study on how artificial intelligence literacy and prompt engineering influence the use of LLMs and GAI in higher education

Supervisor

Prof. Teresa Scantamburlo

Assistant supervisor

Prof. Alessandra Melonio

Graduand

Valentina Rossi
875729

Academic Year

2023/2024

To Severino,
beyond life and its harsh conventions,
to my sister,
the very essence of my existence,
to all who love me,
and whose love has been my courage and strength.

ABSTRACT

The release of ChatGPT by OpenAl on November 2022 marked a significant paradigm shift in the landscape of human-artificial intelligence interaction (HAII); for it signified the accessibility of generative artificial intelligence (GAI) and, in particular, large language models (LLMs) outside the technological élite.

As of August 2024, ChatGPT was reported to have more than 200 million weekly active users, engaging with it to accomplish a wide variety of purposes, including professional or academic activities. To many, ChatGPT represents a polymathic entity, an oracle harnessing the extensive spectrum of human knowledge. Despite the enthusiasm with which this chatbot has been greeted, there is a pressing need to identify the numerous technical, architectural, ethical, and legal constraints associated with its development and usage in higher education.

This study aims to investigate whether possessing adequate literacy on both GAI, with its advantages and limitations, and fundamental prompt engineering techniques, can provide quantitatively and qualitatively improved interactions for university students, together with enhanced awareness of what LLMs can and can not achieve. In this regard, students from various humanities faculties at Ca’ Foscari University of Venice were selected to participate in a literacy workshop and empirical tests in order to contribute towards the development of an effective framework for AI literacy and utilisation within higher education.

ACKNOWLEDGMENTS

My first and deepest appreciation goes to this thesis' supervisor, Professor Teresa Scantamburlo, who has believed in this project since the very beginning, has assisted me wholeheartedly and with dedication, and from whom I have gained extremely valuable insights that I will cherish for my future life journey. Many thanks for demonstrating how the transmission of knowledge extends beyond a mere profession: you inspired me and I will forever be grateful for that.

I would also like to thank the University Ethics Committee for its guidance and suggestions in terms of compliance, so that ethics and legality requirements would not just feature as one of the subjects, but as a rigorous operational paradigm. I am equally grateful to everyone who contributed their suggestions and experience to the realisation of the project.

Special thanks to all the students who, with enthusiasm and commitment, participated in the workshop: their observations and thoughtful involvement were of crucial importance for the success of the pilot study, and could not have been otherwise. Change begins with minor ventures, and they have demonstrated this perfectly with their proactive and curious attitude.

TABLE OF CONTENTS

FOURTH CHAPTER

SECOND PART

INTRODUCTION

The primary objective of this thesis is to present a practical pilot study investigating a pathway for artificial intelligence literacy (AIL) aimed at non-STEM (Science, Technology, Engineering, and Mathematics) undergraduate students at Ca’ Foscari University of Venice. This study seeks to assess the efficacy of an educational initiative centred on the functionality, as well as the ethical and legal implications of generative artificial intelligence (GAI), with a particular focus on large language models (LLMs). Additionally, it introduces students to prompt engineering techniques and patterns, with the aim of enhancing their critical engagement with these technologies. The significance of this research lies in the necessity for non-technical individuals to comprehend and interact critically with AI technologies, which now permeate various aspects of both professional and academic life. AI literacy transcends the technical domain, addressing critical and ethical dimensions, as these technologies raise intricate issues related to privacy, surveillance, algorithmic bias, and decision-making responsibility. By providing students with foundational knowledge of AI, GAI, and prompt engineering, they will be empowered to engage with technology in a conscious and informed manner, which is increasingly vital across all disciplines.

The research was conducted through a three-and-a-half-hour practical workshop which was led by the author, under the supervision of Professor Teresa Scantamburlo. This workshop was complemented by pre- and post-workshop questionnaires and hands-on exercises using ChatGPT. Both to underpin the preparatory work for the AIL workshop and to provide a robust scientific foundation for this thesis, theoretical discussions explored the history of AI – particularly in relation to natural language processing (NLP) – and ethical and legal considerations, as well as the nuances of human-AI interaction (HAII), and prompt engineering techniques and patterns. The thesis further examines AIL, its characteristics and its potential integration into higher education curricula. In addition, the design methodology of the workshop is presented, discussing data analysis and significance of the findings, alongside a detailed examination of ChatGPT's history and functioning. To ensure clarity and coherence in the presentation of the subject matter, the thesis is divided into two main sections: the first addresses theoretical issues, while the second concentrates on the pilot study and ChatGPT.

The first chapter analyses the development and applications of LLMs, with a particular focus on the transformer architecture. This approach has facilitated significant advancements in NLP, by enabling deep learning models to evaluate the relative importance of words (tokens) within a textual sequence, thus improving the handling of long-term linguistic dependencies. The chapter also explores LLMs and GAI through an ethical perspective, particularly underlining the propensity of these models to produce inaccurate or misleading content, commonly referred to as 'hallucinations'. The tensions between the generative capabilities and the legal constraints imposed, for instance, by privacy and intellectual property regulations are examined, highlighting the need for a regulatory framework capable of addressing the challenges posed by these technologies. Additionally, the chapter reflects on the transformative potential of LLMs across various fields, including education, law, and medicine.

The second chapter focuses on HAII, exploring the evolution and implications of GAI technologies in human interaction. It addresses the pivotal role of design in this area and examines the ELIZA effect – a phenomenon wherein users attribute human-like intelligence to machines based on superficial interactions with systems that only appear to exhibit cognitive depth. The chapter underscores the necessity of a Human-Centred approach in artificial intelligence (HCAI), advocating for an ethical and responsible development and integration of AI technologies. They should aim to augment rather than replace human capabilities. Key principles, such as human oversight, transparency, and technical trustworthiness, are explored, concluding with a vision of augmented intelligence that promotes synergistic collaboration between humans and AI.

In the third chapter, the thesis delves into prompt engineering. It involves crafting textual instructions (prompts) designed to elicit coherent, accurate, and high-quality outputs from AI models such as LLMs. Various techniques and patterns are discussed, particularly in the context of education, where they can enhance the creation of learning materials and personalise teaching assistance, thus making learning more engaging and customised.

The fourth chapter reviews the existing literature on AIL and identifies the competencies necessary for navigating a world increasingly shaped by AI and GAI. AI literacy is defined as the ability to understand, utilise, communicate, and collaborate ethically with AI systems. It is vital for fostering critical thinking and facilitating effective interactions. The chapter emphasises the importance of AIL as a core competency, akin to computer literacy, encompassing knowledge of GAI's capabilities and limitations, including hallucinations and the lack of semantic understanding. The relevance of human oversight in decision-making processes is also highlighted.

The fifth chapter concentrates on the purpose and expected outcomes of the empirical study conducted with undergraduate humanities students at Ca' Foscari

University. The pilot study employed a workshop format, comprising two learning modules: one focused on the history, functioning, and ethical and legal implications of GAI and LLMs; the other on prompt engineering techniques and patterns, as well as interaction with ChatGPT. The chapter also provides a detailed overview of OpenAI's history and its technological and ethical challenges. A comprehensive examination of ChatGPT's development, from GPT-1 in 2018 to o1 in 2024, is included, alongside a discussion of the innovations and difficulties associated with the chatbot's evolution.

The sixth chapter outlines the design methodology of the pilot study, covering the workshop, questionnaires, and prompt engineering exercise. An extensive literature review informed the framework. Participants were selected through voluntary application, with 9 students from diverse humanities programmes ultimately participating.

The seventh and final chapter analyses the results of the pilot study through a comparative analysis of pre- and post-workshop data, employing descriptive and inferential statistical methods. The findings indicate that the workshop significantly improved students' understanding of GAI, LLMs and prompt engineering. These results reinforce the importance of integrating AI literacy into curricula to foster informed, ethical engagement with advanced technologies, thereby enhancing students' soft skills, critical thinking, and overall preparedness for the evolving technological landscape.

FIRST PART

FIRST CHAPTER

LARGE LANGUAGE MODELS (LLMs)

"Cogito, ergo sum."
RENÉ DESCARTES, *Discours de la Méthode pour bien conduire sa raison, et chercher la vérité dans les sciences*

"Language is a part of our organism and no less complicated than it."
LUDWIG WITTGENSTEIN, *Tractatus Logico-Philosophicus*

I.1. How machines were taught to 'speak'

Language stands as one of the most defining cognitive faculties distinguishing human beings from other animal species. Since antiquity, philosophy has been deeply engaged in exploring its origins, characteristics, and limitations, as well as its intricate relationship with empirical reality, engaging in discussions still fervent at present. The Greek philosopher Aristotle, in the 4th century B.C., not only underscored the central role of *λόγος*[1] as a medium of communication but also addressed the longstanding question of how it is closely intersected with human cognition. A crucial aspect of the linguistic enquiry emerging from the second text of his *Organon*, the treatise *De interpretatione*, is

[1] *logos*, reasoned discourse.

the concept of language as a formal system governed by specific logical rules. For Aristotle, it serves as an essential tool for understanding sensible reality precisely because its structure mirrors the structure of thought: words constitute the core units of logical reasoning.[2] Continuing along this philosophical trajectory, in the 17th century, Gottfried Wilhelm von Leibniz – renowned for developing the notation for differential and integral calculus still in use today – envisioned a universal artificial mathematical language capable of expressing every facet of human knowledge. In this visionary system, rigorous calculation rules would discern the logical relations between propositions. Leibniz even anticipated the creation of specialised machines capable of executing these calculations.[3]

The analogy between natural language and logic, or algebraic language, constituted a major scientific concern and served as the foundation for Natural Language Processing (NLP), an interdisciplinary field of study that has developed since the 1940s at the intersection of linguistics and computer science. It aims to teach machines to analyse, process and generate human language. Language was being considered a fully-fledged computational problem with the advent of 'calculators': if it could be formalised through logic, then it could consequently be translated into a set of algebraic commands to be executed by a computer. The earliest attempts concern an approach called symbolic artificial intelligence (symbolic AI, or GOFAI, i.e. good old fashioned artificial intelligence), which garnered the most intellectual acclaim until the 1970s. It was based on the assumption of human knowledge being an ordered set of words, combined into sentences (symbols), that could be collected into a series of computational rules, for

[2] The topic is extensively discussed in the second text of his *Organon*, *De interpretatione*, a treatise in which he discussed the structure of prepositions and the various principles of logical inference, thereby laying the foundations for his theory of language as a tool for rational argumentation.

[3] MARTIN DAVIS, *The Universal Computer*, New York, Norton, 2000.

machines to analyse and elaborate on to perform assigned tasks.[4] Symbolic AI systems, such as rule-based machine translation programs, were capable of translating simple sentences. However, they struggled with idiomatic expressions, context, and nuance. For instance, if the system did not understand that *over the moon* is an English expression meaning *being delighted*, it might produce a nonsensical literal translation when converting it to another language. This highlighted a larger problem: the initial reliance on formal logic did not account for the inherent ambiguity of natural language.

Given the close, interconnected correlation between language and cognition, already identified by Plato and Aristotle, and Leibniz, and further deepened during the 20th century by philosophers such as Ferdinand de Saussure[5] or Ludwig Wittgenstein,[6] the possibility that, by learning to process human language, machines would acquire cognitive and intellectual capabilities of their own began to be investigated. This spurred research into 'intelligent machines', and it was also due to the coeval breakthroughs in the field of neuroscience, which demonstrated how the brain was composed of a network of neurons that could be activated by electrical and chemical impulses; as well as the intersection of these concepts with cybernetics[7] and the theory of information.[8]

[4] cf. MELANIE MITCHELL, *Artificial intelligence: a guide for thinking humans*, Pelican, London, 2019.

[5] Ferdinand de Saussure (Geneva, 1857 - Vufflens-le-Château, 1913) is regarded as the father of structural linguistics. In his *Cours de linguistique générale*, he addressed issues such as the difference between *langue*, i.e. the social aspect of language and therefore shared by all speakers, and *parole*, which instead refers to the individual execution. Another fruitful contribution was the introduction of the difference between *signifiant*, the physical form of the sign, and *signifié*, the concept or idea that the sign represents.

[6] Ludwig Josef Johann Wittgenstein (Vienna, 1889 - Cambridge, 1951) is one of the most influential philosophers of the 20th century. In his only work, the *Tractatus Logico-Philosophicus*, he examined the relationship between language and reality. According to him, language is the logical representation of the world and can not exempt itself from what is empirical ('atomic facts'). Whenever language attempts to go beyond sensible reality, then communication fails.

[7] Founded by Norbert Wiener, it explored the systems of control and communication inherent to both living beings and machines. It provided an interdisciplinary approach to mathematics, biology, engineering and philosophy allowing to understand the functioning of complex systems and their interactions with each other.

[8] Proposed by Claude Shannon, it is grounded on the concept of entropy, studying the volume of information held in a data set in order to identify efficient ways of compression and transmission.

The first to comprehensively tackle this issue was the mathematician, computer scientist and logician Alan Mathison Turing, born in 1912 in London and mastermind of the algorithm used to decipher the encrypted messages of the Germans and Italians during the Second World War, significantly contributing to the Allied victory. After the war ended, he relocated to London in order to work at the National Physical Laboratory, where he conceptualised the Automatic Computing Engine (ACE).[9] It is one of the earliest engineering designs for a stored-program computer. Until then, the problem that had prevented machines from being able to learn a set of instructions was the lack of adequate memory to store a significant amount of data.[10] In 1950 Turing published an academic paper shaping the history of artificial intelligence, *Computing Machinery and Intelligence*,[11] in which he speculated on whether machines will ever be able to think. His work marked a turning point in how we conceptualise machine intelligence. Specifically, he began by considering the broad and vague meaning of 'thought', theorising a pragmatic approach to assess whether a machine is capable of cognition. This test, which he named *The Imitation Game*, has conditioned the common idea of intelligent machines up to contemporary times. Turing hypothesised that if a human judge, during a conversation, failed to distinguish between the answers formulated by another human and a computer, then the latter could have been assumed to possess human-level intelligence. Nevertheless, this method gives more importance to imitating surface-level behaviours rather than possessing authentic cognitive comprehension.

> Consider first the more accurate form of the question. I believe that in about fifty

[9] *Alan Turing*, in *Biography*, https://www.biography.com/scientists/alan-turing, accessed on 20.02.2024.

[10] Early forms of memory consisted mainly of mechanical or electromechanical devices.

[11] ALAN MATHISON TURING, *Computing Machinery and Intelligence*, in «Mind», volume LIX, issue 236, 1 Oct. 1950, pp. 433-460.

> years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, 'Can machines think!' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.[12]

A few years afterwards, in 1956, the expression 'artificial intelligence' was coined at the emblematic Dartmouth Workshop.[13] The premises of the proposal submitted by the computer scientists John McCarthy, Marvin Lee Minsky, Nathaniel Rochester and Claude Shannon were coherent with the insight to inject cognition into computers through language. «[A]n attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves».[14] This workshop set ambitious goals for the field and laid the groundwork for future AI research, but above all, it crystallised the still-cherished objective of equalling human intellect through machines. «For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving».[15]

The initial approaches to NLP took place in the context of Machine Translation (MT) between different languages, e.g. from English into Russian, as evidenced at the

[12] *Ibidem.*

[13] Cf. JOHN MCCARTHY, ET AL., *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 31 Aug. 1955, https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, accessed on 20.02.2024.

[14] *Ibidem.*

[15] *Ibidem.*

IBM-Georgetown demonstration in 1954.[16] In 1957, the American linguist Avram Noam Chomsky,[17] published a pivotal essay: *Syntactic Structures*,[18] shifting the focus from the description of specific languages to broader research into the common structure of human natural language, which became the fundaments for instructions that machines could use.

Between 1964 and 1966, the computer scientist Joseph Weizenbaum wrote the first chatbot in history, ELIZA,[19] a simulation of a Rogerian psychotherapist,[20] at the Massachusetts Institute of Technology (MIT). The experiment raised questions about the definition of artificial intelligence and, in particular, the Turing test, which until then was the paradigm of reference. Although the chatbot fairly simulated human-like conversational skills, that did not mean it possessed similar cognitive abilities or any depth. «The NLP ambitions for ELIZA were avowedly modest, but even assessed within these parameters the program had many limitations. It was (and is) easy to 'trick', it can be made to loop recursively, and is quickly 'persuaded' to come out with nonsensical answers».[21] Notwithstanding these clear constraints, its success encouraged the emergence of numerous studies in the field of human-computer interaction (HCI), even to the point of naming as the 'Eliza effect' the phenomenon «defined as the susceptibility of people to read far more understanding than is warranted into strings of symbols –

[16] KAREN SPARCK JONES, *Natural Language Processing: A Historical Review*, in *Current Issues in Computational Linguistics: In Honour of Don Walker*, NICOLETTA CALZOLARI, MARTHA PALMER AND ANTONIO ZAMPOLLI (edited by), Springer, Dordrecht, 1994.

[17] (Philadelphia, 1928) largely influenced scientific development in the field of linguistics through his theory of generative grammar and linguistic innatism: he argued that language is an inherent capacity of the human being as self.

[18] NOAM CHOMSKY, *Syntactic structures*, Berlin, New York, De Gruyter Mouton, 2002.

[19] JOSEPH WEIZENBAUM, *ELIZA - a computer program for the study of natural language communication between man and machine*, in «Communications of the ACM», volume 9, number 1, pp. 36-45.

[20] An approach characterised by empathic listening and reflection on the words chosen by the patient.

[21] CAROLINE BASSET, *The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present*, in «AI & Society», volume 34, pp. 803-812.

especially words – strung together by computers».[22] These concepts will be explored in more detail in the second chapter.

Throughout the history of artificial intelligence, however, there have been as many springs as winters: in the 1970s, the promises gathered from that Dartmouth workshop onwards, which aspired to build an intelligent machine within the period of a generation, collapsed under the weight of their expectations.[23] In particular, researchers realised that producing a list of instructions was not sufficient for machines to conquer the ability to contextualise what they were processing, i.e. to achieve that 'common sense' of the world which is predominantly implicit in human cognition. A more complex and multifaceted approach was needed, and this was machine learning: «the process of computers improving their own ability to carry out tasks by analysing new data, without a human needing to give instructions in the form of a program».[24]

While Marvin Lee Minsky and John McCarthy were lavishly financed at the AI laboratories they founded, short-sighted about the prospects of symbolic AI, psychologist Frank Rosenblatt tried to solve the problem by taking inspiration directly from neuroscience. In 1958, he theorised the perceptron (Figure 1),[25] a sort of calculational neuron (and the ancestor of current artificial neurons). A neuron is a cell within the human brain that receives and exchanges electrical or chemical signals with the other neurons to whom it is connected: its activation is determined by the sum of the inputs received

[22] DOUGLAS RICHARD HOFSTADTER, FLUID ANALOGIES RESEARCH GROUP, *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanism of Thought*, New York, Basic Books, 1995.

[23] JAMES LIGHTHILL, *Artificial Intelligence: A General Survey*, in «Artificial Intelligence: a paper symposium», Science Research Council, London, 1973.

[24] *Machine Learning*, in *Cambridge Dictionary*, https://dictionary.cambridge.org/dictionary/english/machine-learning, accessed on 15.03.2024.

[25] FRANK ROSENBLATT, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*, in «Psychological Review», volume 65, number 6, 1950, pp. 386-408.

(having different levels of strength according to the weight of the connection) which has to reach a specific threshold in order to trigger its activation.

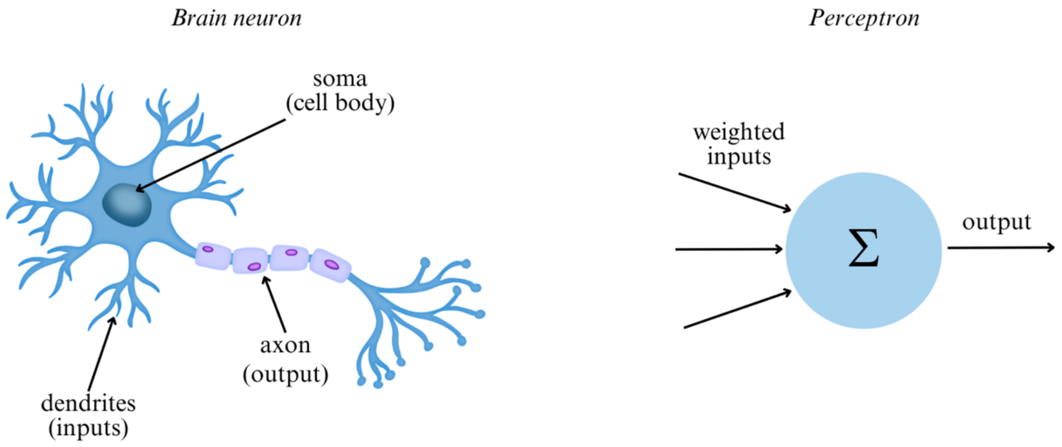


FIGURE 1, *Analogies between a brain neuron and Rosenblatt's perceptron.*

Analogously, the perceptron receives numeric inputs ($x_1, x_2, \dots x_n$) and combines them through a weighted sum, in which each input x_i is multiplied by a corresponding weight w_i. This weighted sum is then passed through an activation function, usually a threshold function, which determines a binary output (0 or 1).

$$y = sign\left(\sum_{i=1}^{n} w_i x_i\right)$$

y represents the binary output predicted by the perceptron (0 or 1)

$sign$ is a function which returns +1 if its argument is positive and -1 if negative.

$\sum_{i=1}^{n} w_i x_i$ is the weighted sum of the inputs, where:

w_i represents the weights associated with the inputs

x_i represents the individual inputs

n is the total number of inputs.

This is achieved through training: the perceptron network is given inputs by which to take a guess, subsequently the programmer provides a signal telling the system how far the output is from the correct one, and the perceptron network can thus adjust its weights and thresholds automatically.[26] Still, the perceptron being a linear classifier can only divide data that are linearly separable. A classic example highlighting this limitation is the XOR problem, a binary logic function that returns true (1) if and only if one of the two inputs is true (1), but not both.

The research – although promising – suffered a setback in 1969, when Marvin Lee Minsky and Seymour Aubrey Paupert published the book *Perceptrons: an Introduction to Computational Geometry*,[27] underlining its limitations, especially in terms of scale-up, and considering Rosenblatt's contributions fundamentally sterile. Yet, they also suggested how adding a layer could significantly improve performance: this type of architecture is called multi-layer perceptron (MLP).

Thanks to small groups of researchers (particularly in the field of psychology), Rosenblatt's studies continued to be improved. In 1987, a group at the University of California wrote a treatise, *Parallel Distributed Processing*,[28] in which more depth was given to so-called 'connectionist networks' (artificial neural networks, ANN) (Figure 2). This provided a driving impulse to subsymbolic AI, where knowledge resides in the weighted connections between the different neurons, and not in a previously provided set of rules as for symbolic AI.

[26] Cf. M. Mitchell, *Artificial intelligence: a guide for thinking humans*, cit.

[27] Marvin Lee Minsky, Seymour Aubrey Paupert, *Perceptrons: an Introduction to Computational Geometry*, MIT Press, Cambridge (Massachusetts), 1969.

[28] cf. David Emerett Rumelhart, James Lloyd McClelland, PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, The MIT Press, Cambridge (Massachusetts), 1987.

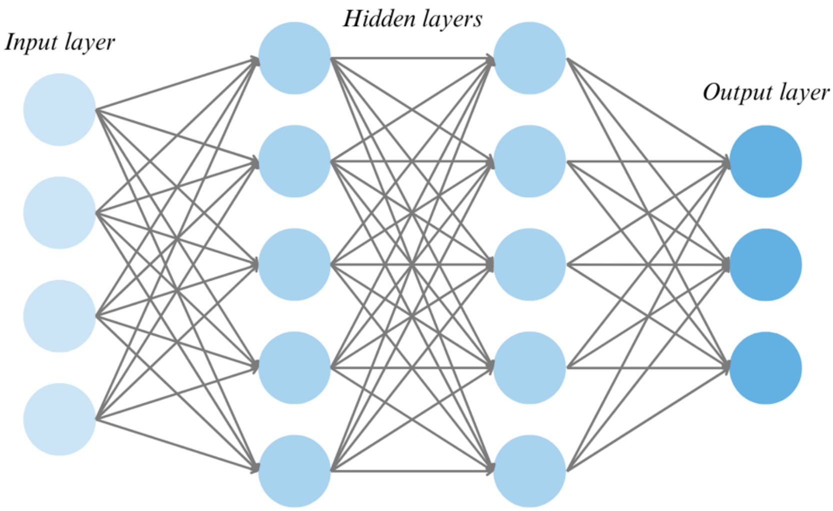


FIGURE 2, *Artificial neural network.*

Behind this turnaround was back-propagation, an algorithm enabling artificial neural networks to autonomously adjust the weights of neurons' connections by learning from the data iteratively.[29] It involves the loss function $\mathrm{L}(\mathrm{y}, \hat{y})$ which quantifies the difference between the network's predicted output y and the actual target $\hat{y}$. To minimise this loss, the chain rule of calculus is applied to compute the gradient of the loss function concerning each weight in the network. By following the direction reducing the loss (gradient descent), the network iteratively adjusts its weights to improve performance.

$$w_i \leftarrow w_i - \eta \cdot \frac{\delta \mathrm{L}(\mathrm{y}, \hat{y})}{\delta w_i}$$

w_i represents the weight associated with the input x_i

η is the learning rate

$\frac{\delta \mathrm{L}(\mathrm{y}, \hat{y})}{\delta w_i}$ is the gradient of the loss function $\mathrm{L}(\mathrm{y}, \hat{y})$ with respect to weight w_i, this represents how much the loss changes as the weight w_i changes.

[29] Cf. M. MITCHELL, *Artificial intelligence: a guide for thinking humans*, cit.

However, between the 1980s and 1990s, AI experienced another winter, as a consequence of an economic bubble that saw numerous companies go bankrupt. The difficulty in raising funds for further research did not prevent the advancement of ANNs, also thanks to statistics and probabilistic theory. The advantages in the field of NLP were particularly stimulating. Frederick Jelinek,[30] an information-theoretic linguist, was even attributed the emblematic citation «every time I fire a linguist, the performance of the speech recognizer goes up»[31] to highlight the effectiveness of machine learning techniques in comparison with subsymbolic AI.

In 1997 an advancement in the field of NLP occurred thanks to recurrent neural networks (RNNs) (Figure 3), «loosely inspired by [the] sequential process of reading a sentence and creating a representation of it in the form of neural activations».[32] RNNs are designed to handle sequential data, which is crucial for linguistic tasks, as the order of words significantly impacts meaning. In contrast to traditional feedforward neural networks where inputs and outputs are independent, RNNs exploit a 'memory' mechanism. Their «hidden units have additional "recurrent" connections; each hidden unit has a connection to itself and to the other hidden unit».[33] Therefore, each neuron calculates its activation by considering both the weighted inputs and the activations of the neurons of the previous time step. In other words, every word in a sentence is not only processed separately but also related to the words that preceded it.

[30] (Kladno, 1932 – Baltimore, 2010) was a researcher at IBM, where he contributed to computer speech recognition and machine translation.
[31] Cited in NELLO CRISTIANINI, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano,* Il Mulino, Bologna, 2023.
[32] M. MITCHELL, *Artificial intelligence: a guide for thinking humans*, cit.
[33] EAD., p.

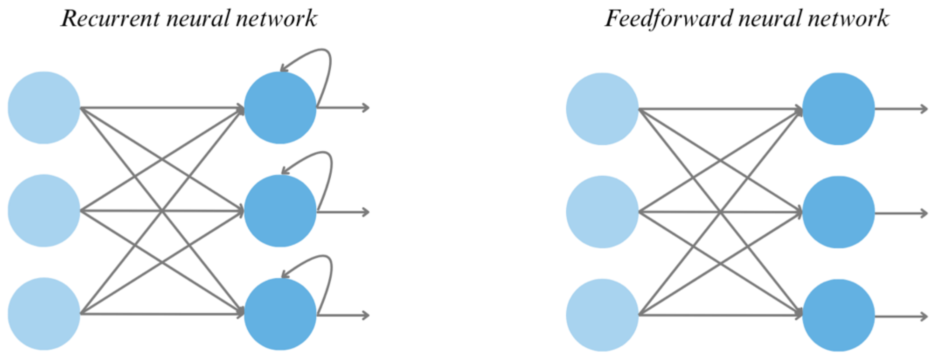


FIGURE 3, *Differences between a neural network and a feedforward neural network.*

In a linguistic context, this approach allows for more accurate outputs, as it does not consider lemmas individually, but it enables the network to retain an awareness of the sentence structure and meaning as it unfolds.

As previously mentioned, the inputs of the neurons that compose a neural network have to be numerical values in order to be summed up and multiplied by their weights. Clearly, this raises a problem of a structural sort, for words have to be converted into numbers. Hence, from distributional semantics, based on the quote «you shall know a word by the company it keeps»,[34] an algorithm was developed in the 2010s. It aimed at encoding words by capturing the semantic relationship linking them to others. By converting them into geometric points within a three-dimensional semantic space, each word can be identified by its coordinates – its position in the *x*, *y* and *z* axis (e.g. *word2vec*) (Figure 4).[35] This approach is usually referred to in NLP as a vector of words, or word embedding, and it currently constitutes the foundations of the architecture behind most language models.[36]

[34] Cited in EAD.
[35] XIN RONG, *word2vec Parameter Learning Explained*, in *ArXiv*, 11 Nov. 2014, https://doi.org/10.48550/arXiv.1411.2738, accessed on 27.02.2024.
[36] cf. TIANYU WU, ET AL., *A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development*, in «IEEE/CAA Journal of Automatica Sinica», vol. 10, no. 5, 2023, p. 1124.

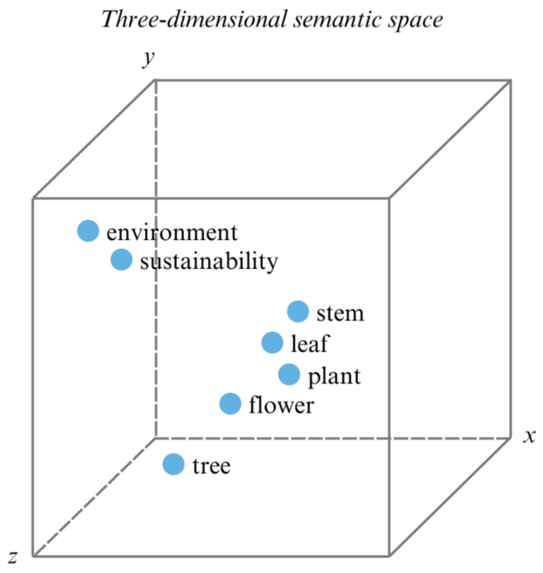


FIGURE 4, *Representation of a hypothetical three-dimensional semantic space in which words are represented as points.*

Deep learning[37] and word embeddings provided an invaluable impulse to the field of NLP, and – although today more sophisticated evolutions of these architectures are relied upon, to be discussed in more detail – they ushered in a breakthrough. During the 2010s it was witnessed the birth of voice assistants such as Siri developed by Apple Inc., or Alexa integrated by Amazon.com, Inc., as well as machine translation programmes such as Google Translate by Google LLC,[38] and a wide range of natural language processing programmes.

The limitations of RNNs, even in their more advanced forms such as LSTMs (long-short term memory),[39] drove research towards finding new algorithms capable of handling the complexities of natural language more effectively. In 2017, the paper

[37] A branch of machine learning that employs multi-layered neural networks, known as deep neural networks, to replicate the intricate decision-making abilities of the human brain.
[38] Only in 2016, however, did it change its architecture to neural machine translation (NMT).
[39] They are designed to handle long-term dependencies within sequential data, through internal mechanisms controlling the flow of information through specific gates.

Attention is all you need issued by Ashish Vaswani, et al.,[40] all associate researchers at the Google Brain team (dedicated to leading-edge research in machine learning), ushered in the transformer architecture, which revolutionised the field of NLP and laid the foundations for the development of large language models (LLMs). Transformers constitute a subversive breakthrough as they move forward from the sequential nature typical of RNNs (which process inputs one at a time in order, taking into account the hidden state deriving from previous inputs). The objective of their attention mechanism is to determine the relative importance of each word about the others, capturing their contextual relationships regardless of the distance within the sequence. Each word is represented by three vectors: *query* (*Q*), *key* (*K*), and *value* (*V*). The scalar product between *query* and *key* calculates how 'similar', or relevant, it is a word with regards to the others, producing scores that are normalised through a SoftMax function (converting these values into a probability distribution). The sum of the probabilities for each word must be 1. The resulting probabilities are utilised in a weighted sum, determining the *value* vectors of all words, and used to weight the *value* vectors. «[This] mechanism allows every word to attend to all previous words or every word except the target, allowing the model to efficiently capture long-range dependencies without the expensive recurrent computation in LSTMs».[41]

A key aspect of the transformer is multi-head attention. Rather than applying a singular attention task, the transformer splits the process into multiple 'heads', each analysing the sequence independently. The outputs obtained from each head are then

[40] ASHISH VASWANI, ET AL., *Attention is all you need*, paper presented at the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017.

[41] BONAN MIN, ET AL., *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*, in *ArXiv*, 1 Nov. 2021, https://doi.org/10.48550/arXiv.2111.01243, p. 3, accessed on 15.03.2024.

linked and combined, allowing the model to capture different perspectives and relationships between words in a richer and more detailed manner. The main architecture of a transformer (Figure 5) essentially consists of an encoder and a decoder, each comprising an attention module and a feedforward neural network.

Transformer architecture

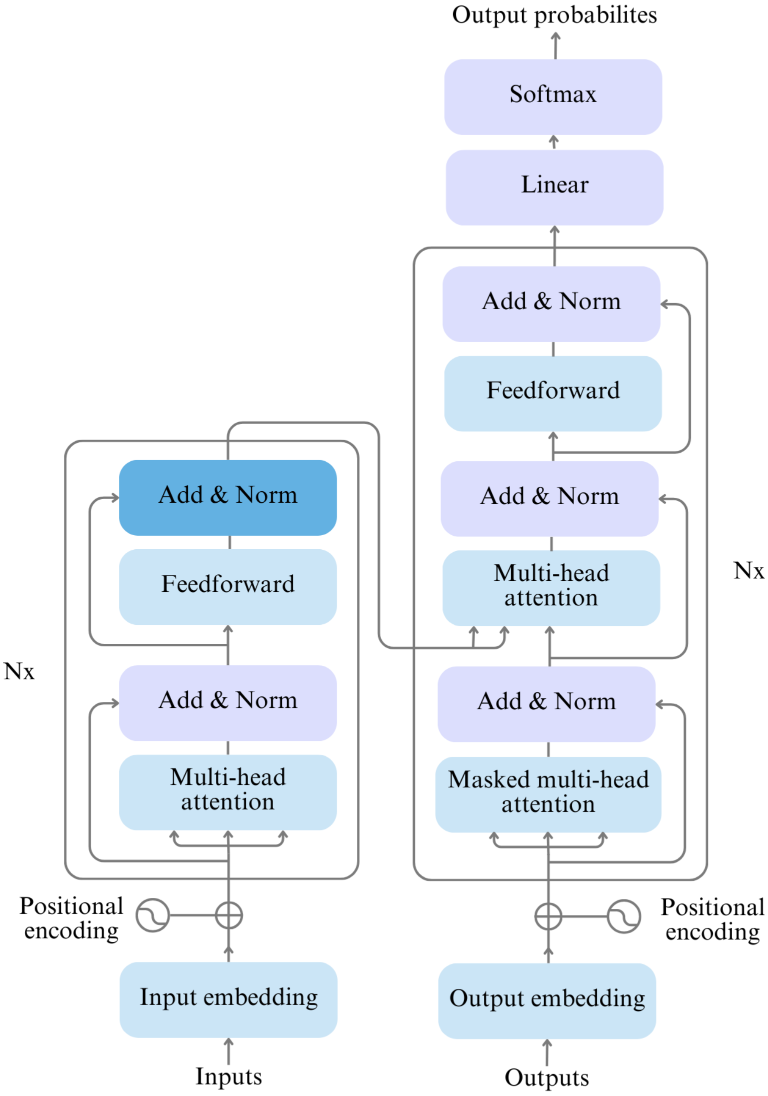


FIGURE 5, *Traditional transformer architecture.*

The encoder (left) receives the textual sequence as input, and each token (word or sub-words) is first transformed into a vector in a continuous space through embedding, as mentioned earlier. Since transformers do not possess an intrinsic sequential structure (like RNNs), information about the position of tokens within the sequence must be explicitly added through positional encodings, i.e. vectors added to incorporate this additional information. The encoder is composed of a stack of N identical layers (usually 6, 12, 24, or 48 for very large AI models such as GPT-4), and each layer by two sublayers. The first implements the multi-head attention mechanism described above, and the second is a feedforward network, which applies linear transformations (such as ReLU). All layers apply the same linear transformation on all individual input tokens, but each layer uses different weights and biases. To improve the stability of training and facilitate the flow of information, each layer has a normalisation layer and skip connections. On the other hand, the decoder (right) is responsible for transforming the representation generated by the encoder into an output sequence (in the context of machine translation, for instance, it is the translated word sequence), taking advantage of an autoregressive mechanism. The output of the previous step will then be taken as input to the decoder in the following step. This is achieved by a masked self-attention mechanism preventing the model from seeing the subsequent words during the generation of the output. As with the encoder, each layer of the decoder includes a feedforward network, followed by a normalisation layer and skip connections. At the end of the layers stack, the output is a vector of representations for each token within the textual sequence. For language generation tasks, the last layer is a SoftMax function producing a probability distribution over a vocabulary of possible successive tokens (thus predicting the probability that word ω_{t-1} will be followed by word ω_t).

This new paradigm has paved the way for the development of LLMs, large-scale language models that leverage the capabilities of deep learning to process huge amounts of textual data to generate coherent and sophisticated content. In addition, it also allowed (together with increased computational capabilities with more powerful and efficient hardware being developed,[42] and greater data availability thanks to the advent of the Internet)[43] the exponential growth of generative artificial intelligence (GAI). Differently from traditional AI algorithms, which were limited to the parsing and pattern recognition within a dataset, generative AI is specifically designed to be capable of generating content, be it text, images, videos, or songs, building on vast training datasets.

I.2. LLMs: definition, functioning, and limitations

There is no universally accepted definition for what constitutes an LLM nor is there a set of parameters and characteristics that can be used to categorise generative AI models within a specific range.

> Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.[44]
>
> A complex mathematical representation of language that is based on very large amounts of data and allows computers to produce language that seems similar to

[42] JARED KAPLAN, ET AL., *Scaling Laws for Neural Language Models*, in *ArXiv*, 23 Jan. 2020, https://doi.org/10.48550/arXiv.2001.08361, accessed on 15.03.2024.

[43] PAOLO VILLALOBOS, ET AL., *Will we run out of data? Limits of LLM scaling based on human-generated data*, in *ArXiv*, 26 Oct. 2022, https://doi.org/10.48550/arXiv.2211.04325, accessed on 15.03.2024.

[44] *What are large language models (LLMs)?*, in *IBM*, https://www.ibm.com/topics/large-language-models, accessed on 17.03.2024.

what a human might say.[45]

Yet, one of the most concise and pertinent definitions seems to be the one provided by Google in one of its open-source training courses:

> Large language models refer to **large**, **general-purpose** language models that can be **pre-trained** and then **fine-tuned** for specific purpose.[46]

Large

In this instance the term carries a double meaning, referring both to the impressive size of the training dataset and to the model's vast number of parameters. LLMs are pre-trained on huge volumes of textual data, frequently on the terabyte scale, and include a wide range of texts such as novels, scholarly articles, blogs, conversations, and so forth. GPT-4,[47] the most sophisticated model developed by OpenAI to date (which will be discussed in greater detail in the second part of this thesis), is believed to have been trained on 13 trillion tokens (approximately 10 trillion words), exploiting data from sources such as Common Crawl,[48] an archive of copies of webpages collected by a non-profit organisation bearing the same name.[49] However, these models are often trained also on materials protected by intellectual property (IP) rights, and due to the opaque nature of their functioning[50] and the non-disclosure of details regarding the training data by

[45] *Large language model*, in Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/large-language-model, accessed on 14.03.2024.
[46] Cited in *Introduction to Large Language Models*, in *Google Cloud Skills Boost*, https://www.cloudskillsboost.google/course_templates/539, accessed on 14.03.2024.
[47] JOSH ACHIAM, ET AL., *GPT-4 Technical Report*, in *ArXiv*, 24 May 2019, https://doi.org/10.48550/arXiv.2303.08774, accessed on 15.03.2024.
[48] https://commoncrawl.org/
[49] DYLAN PATEL, AND GERALD WONG, *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*, https://www.semianalysis.com/p/gpt-4-architecture-infrastructure, accessed on 15.02.2024.
[50] They are often referred to as 'black boxes'.

providers, it is complex to determine whether this is the case. This chapter's subsequent section (I.4) will delve further into these ethical and legal constraints.

The term 'large' also refers to the size of the model in terms of parameters: i.e. numerical values learnt by the neural network during the training phase, determining how it processes information and, as a result, its overall performance. For instance, GPT-4 has approximately 1 trillion parameters across 120 layers.[51] Due to their massive size, LLMs typically outperform their predecessors, pre-trained language models (PLMs), by exhibiting emergent abilities such as in-context learning, instruction following, multi-step reasoning, in addition to the possibility of efficiently interacting with users.[52] «The emergent abilities demonstrated by LLMs make it possible to build general-purpose AI agents based on LLMs. While LLMs are trained to produce responses in static settings, AI agents need to take actions to interact with dynamic environment».[53] This dimension advantage also allows them to generalise effectively across a wide range of tasks without being explicitly trained on each one. Despite the advantages, recently, smaller models have also been introduced, driven by the high computational and energetic costs associated with the training and deployment of larger models. Furthermore, smaller models offer faster inference times, making them more suitable for some applications.[54]

[51] *Ibidem*.
[52] SHERVIN MINAEE, ET AL., *Large Language Models: A Survey*, in *ArXiv*, 20 Feb. 2024, https://doi.org/10.48550/arXiv.2402.06196, accessed on 20.07.2024.
[53] *Ibidem.*
[54] An example of such a smaller model is Llama 3 8B, released by Meta, where "8B" refers to the model's 8 billion parameters – a significantly smaller number compared to the parameters of GPT-4 mentioned earlier.
J. WEI, ET AL., *Emergent Abilities of Large Language Models*, in *ArXiv*, 26 Oct. 2022, https://doi.org/10.48550/arXiv.2206.07682, accessed on 20.05.2024.

General-purpose

LLMs are not designed to address a specific problem, such as language translation, rather they are versatile tools that can be applied to a wide range of linguistic tasks, including text completion, creative writing, question answering, sentiment analysis, code generation, and more. The text generation process is based on the prediction of the following word (or token) within a textual sequence, taking into consideration the preceding ones. In principle, the model could choose the highest probability word (greedy sampling), and this would ensure the generation of a text strictly following the linguistic structures learnt during the pre-training phase. However, this approach would lead to repetitive, foreseeable texts, lacking creativity.

To enhance the diversity and originality of a generated text, several techniques are employed: one of the best known is top-k sampling. Instead of simply selecting the word with the highest probability, the model considers the k most probable words (where k is a predefined number) to choose from. This technique allows to explore linguistic choices that are not necessarily the most probable, but may nevertheless be consistent and interesting. This prediction is formalised through the conditional probability of the following word ω_t, given the context of the preceding words $\omega_1, \omega_2, \ldots, \omega_{t-1}$ within the textual sequence. The function calculating this probability is often implemented as a SoftMax function applied to the raw values (logit) z_t the transformer returns as output of the last layer.

$$P(\omega_t | \omega_1, \omega_2, \ldots, \omega_{t-1}) = softmax(z_t)$$

$P(\omega_t | \omega_1, \omega_2, \ldots, \omega_{t-1})$ expresses the probability that the word ω_t will be next in the textual sequence, given the preceding words. The LLM has learnt in the pre-training phase how words combine with each other in various contexts

$softmax$ is the probability calculation function

z_t are real numbers (positive or negative) representing the non-normalised evidence for each possible word that could follow in the text sequence.

Once these recalibrated probabilities have been calculated, the final word is chosen through a weighted sampling: this means that the word with a higher recalibrated probability has a higher probability of being chosen, yet all words in the subset V_k still have a chance to be selected. This method balances consistency with creativity. The most probable words are always more likely to be selected, but other words are not completely excluded. Therefore, despite the selection process being stochastic (i.e. based on a probability) it is not completely randomised.

There are alternative techniques, such as top-p sampling (also known as nucleus sampling), which instead of just considering the k most probable words, selects a dynamic subset of words V_p based on cumulative probability, only comprising those whose probabilities cumulate up to a predetermined threshold p (where $0 < p \leq 1$). A weighted sampling is then performed, similar to top-k, but with the added advantage of the number of words considered varying according to the probability distribution. This approach allows for greater adaptability to the context and greater overall efficiency.

Pre-trained

As above mentioned, the training phase is crucial for ensuring the model is able to reach optimal performance. Effective pre-training necessitates carefully designed model architectures, acceleration strategies, and optimisation techniques.[55] During this specific

[55] JACOB DEVLIN, ET AL., *BERT: Pre-training of Deep Bidirectional Transformers for Language*

stage, the LLM is exposed to a vast dataset comprising unlabelled inputs: this type of training is known as self-supervised learning. In self-supervised learning, the model generates its own labels from the data, typically by predicting parts of the input based on other parts, allowing it to learn autonomously. Unlike supervised learning, where the model learns from a dataset with explicitly assigned labels, self-supervised learning enables the model to uncover patterns, relationships, and hidden structures within the data without needing manually labelled examples. One of the most important aspects is related to data collection and processing, as the principle of GIGO (*garbage in, garbage out*) applies here. Poor-quality inputs leads to poor-quality outputs. Moreover, machine learning often entails the risk of amplifying the biases present in the training data. Specifically, the vectorial representation of the tokens, i.e. word embeddings, does indeed allow the model to learn the relationships between words and their respective co-occurrences, but it simultaneously triggers the acquisition of the biases present in the data, which are then directly transferred to the vectorial representations themselves. For instance, if, within the training records, the word 'computer programmer' is predominantly found in connection with words and pronouns pertaining to a male context, [56] the model would unconsciously absorb this gender bias and equally unconsciously could then reflect it back when generating an output.[57] This phenomenon does not only occur with gender biases, but with any other type of prejudice: social,

Understanding, in *ArXiv*, 24 May 2019, https://doi.org/10.48550/arXiv.1810.04805, accessed on 15.03.2024.

[56] TOLGA BOLUKBASI, ET AL., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, in *ArXiv*, 21 Jul. 2016, https://doi.org/10.48550/arXiv.1607.06520, accessed on 29.03.2024.

[57] UNESCO, *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. in *UNESCO*, 7 Mar. 2024, https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes, accessed on 20.04.2024.

religious, racial, and so on:[58] this ethical dilemma will be addressed in section I.4 of this chapter.

After gathering a wealth of textual data, proceeding with the removal of noisy, redundant, unnecessary, and potentially harmful items is indispensable, as these can negatively affect the skill and behaviour of the model. A typical data pre-processing workflow includes several key steps:[59]

- Data cleaning. Poor-quality or unreliable data must be eliminated, ensuring that the pre-training corpus consists of high value texts. Targeted criteria, such as linguistic accuracy, correct formatting and content relevance, are employed to retain only useful data.
- Data filtering. Most common data filtering techniques include noise removal, i.e. data that might affect the LLM's ability to generalise, handling outliers and ambiguities that could confuse or disproportionately affect the model, balancing the distribution of data to minimise bias, as well as text pre-processing, by standardising and cleaning the textual data.
- De-duplication. The presence of redundant data can increase biases and unnecessarily burden the size of the corpus. This process removes repeated materials, resulting in a cleaner and lighter dataset.
- Privacy. To protect sensitive information (or, for instance, IP protected materials) privacy redaction techniques are applied, removing or obscuring any type of personally identifiable information and confidential data.[60]

[58] RISHI BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, in *ArXiv*, 22 Jul. 2022, https://doi.org/10.48550/arXiv.2108.07258, accessed on 17.03.2024.
[59] S. MINAEE, ET AL., *Large Language Models: A Survey*, cit.
[59] R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit.
[60] In Italy, as well as throughout the European Union, the processing of personal and sensitive data is

- Tokenisation. This technique consists of decomposing the textual sequence into smaller units, called tokens (words, sub-words or even specific characters, depending on the technique adopted). This step is necessary for language modelling.

In order to avoid overfitting (i.e. when a statistical model adapts too closely to the training data, without being able to generalise), regularisation techniques such as dropout are applied. Dropout randomly deactivates a fraction of units (neurons) within the neural network during the pre-training phase, and the deactivated units do not participate in the forward pass and backpropagation process at a given stage. It is efficient because, by randomly deactivating certain portions of the network, it prevents the model from becoming overly dependent on specific neurons or connections.

Successful performance of an LLM is heavily dependent on the training phase, which is in turn highly correlated with the computational capabilities of the hardware on which these models are being deployed. The hardware utilised is typically composed of graphics processing units (GPUs), tensor processing units (TPUs), and high-performance central processing units (CPUs). GPUs are optimised to perform large-scale parallel computing operations. Each unit contains thousands of processor cores, each of which can perform several computing operations simultaneously. This is the reason why GPUs are therefore optimal for the set of matrix multiplication operations required by transformers. TPUs are hardware accelerators specifically designed by Google to improve

regulated by the GDPR, a regulation in force since 2016 and with which many providers of LLMs have often experienced discrepancies. The Italian Data Protection Authority (Garante della Privacy), the designated Italian authority, has charged OpenAI with two violations in March 2023 (which led to a suspension of its services in Italy for a few weeks) and January 2024.

the efficiency of deep learning model training.[61] They are optimised for performing tensor operations, which are multi-dimensional data structures utilised in all major deep learning frameworks such as TensorFlow and PyTorch. Lastly, CPUs (although less optimised for massively parallel computing than GPUs and TPUs) handle operations at a more general level, such as loading data, managing interfaces with memory, and so on. Advanced memory systems, such as random access memory (RAM) and high-bandwidth memory (HBM), are also included in the hardware architecture and are crucial for storing the heavy datasets and huge model parameters during training. High-speed interconnection between processing units, such as NVLink in NVIDIA's GPUs, enables fast communication between GPUs, further improving the efficiency of distributed training across multiple processing units, and has propelled the company to become the most valuable public company in such a short time since the generative AI industry burst onto the scene.[62] Significantly, this type of training is environmentally extremely wasteful, due to the enormous amount of energy required to power this high-performance hardware. Such massive energy consumption results in significant CO_2 emissions (in the order of thousands of tonnes),[63] and will be discussed more specifically in subsection I.2.2 of this section. By the end of the pre-training phase, the model has developed a general understanding of natural language – encompassing vocabulary, grammar, and logical relationships – though it is not yet fine-tuned for more specific tasks.[64]

[61] NORMAN PAUL JOUPPI, ET AL., *In-Datacenter Performance Analysis of a Tensor Processing Unit*, paper presented at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, 2017, https://doi.org/10.48550/arXiv.1704.04760, accessed on 20.07.2024.

[62] TRIPP MICKLE, JOE RENNISON, *Nvidia Becomes Most Valuable Public Company, Topping Microsoft*, in in «The New York Times», 18 Jun. 2024, https://www.nytimes.com/2024/06/18/technology/nvidia-most-valuable-company.html, accessed on 20.07.2024.

[63] DAVID PATTERSON, ET AL., *Carbon Emissions and Large Neural Network Training*, in *ArXiv*, 23 Apr. 2021, https://doi.org/10.48550/arXiv.2104.10350, accessed on 20.07.2024.

[64] B. MIN, ET AL., *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*, cit.

Fine-tuned

Fine-tuning is a machine learning technique in which a pre-trained language model undergoes additional training on a specific dataset, in order to specialise for a particular domain or task. This process exploits the general knowledge acquired during the pre-training phase, incorporating the specialised information learnt during this second stage. The dataset is typically smaller, more specific to avoid overwriting the general data, and labelled. It contains examples relevant to the desired task; for instance, an LLM can be fine-tuned to generate answers for a medical chatbot using conversations between doctors and patients.

The data needs to be pre-processed. Besides traditional techniques such as tokenisation and segmentation, sometimes data augmentation is employed to expand this second dataset. The sequence-to-sequence (seq2seq) architecture is a popular technique for segmentation in fine-tuning. It is particularly useful for tasks where an input sequence needs to be turned into a specific output sequence, such as machine translation or question answering. The seq2seq architecture, too, consists of an encoder and a decoder. The encoder processes the input sequence and converts it into a numerical representation (vector), while the decoder utilises this representation to generate the output sequence. During fine-tuning, the model's parameters are adjusted through backpropagation to reduce the difference between its predictions and the desired behaviour. However, some of the model's layers are commonly frozen to avoid overfitting and reduce computational costs.

A lighter technique to tune LLMs to specific tasks or domains is prompt tuning. Rather than updating all (or a significant part) of the model parameters, this strategy allows to focus on optimising a small set of additional parameters, called prompt embeddings. These are not part of the pre-trained model but are concatenated with the original inputs as additional vectors, thus influencing the output. Since only a small set of parameters is optimised, prompt tuning is much less computationally burdensome, however, it fails to capture very domain-specific complexities that require deeper modifications to the model parameters.

Another step to improve the performance of fine-tuned LMMs is the adoption of reinforcement learning from human feedback (RLHF),[65] which combines supervised learning and reinforcement learning techniques to further refine the model based on human feedback. Specifically, reinforcement learning is a machine learning paradigm in which an agent learns to perform a certain task through several trial-and-error interactions, to maximise rewards depending on performance. RLHF is aimed at reaching both optimal statistical performance and coherent and effective behaviour in real-world contexts. When it comes to generative tasks, it is crucial to satisfy requirements of relevance, adequacy, and coherence with human expectations. This approach consists of two phases. Firstly, a reward model is trained on labelled data, where the main model's responses are evaluated considering their quality: each pair of answers is marked with the preferred one. This reward model learns to predict a scale of rewards based on human preferences. Subsequently, the main model is optimised through reinforcement learning, leveraging techniques such as proximal policy optimisation (PPO) to maximise the predicted reward, as it efficiently balances exploration and exploitation. Human feedback is consistently incorporated. As the model generates new outputs, they are re-labelled and used to update both the reward model and the policy of the main model. A critical aspect

[65] LONG OUYANG, ET AL., *Training language models to follow instructions with human feedback*, in *ArXiv*, 4 Mar. 2022, https://doi.org/10.48550/arXiv.2203.02155, accessed on 20.03.2024.

of RLHF is finding a balance in the rewards to prevent unintended behaviours, such as flat or overly compliant answers.

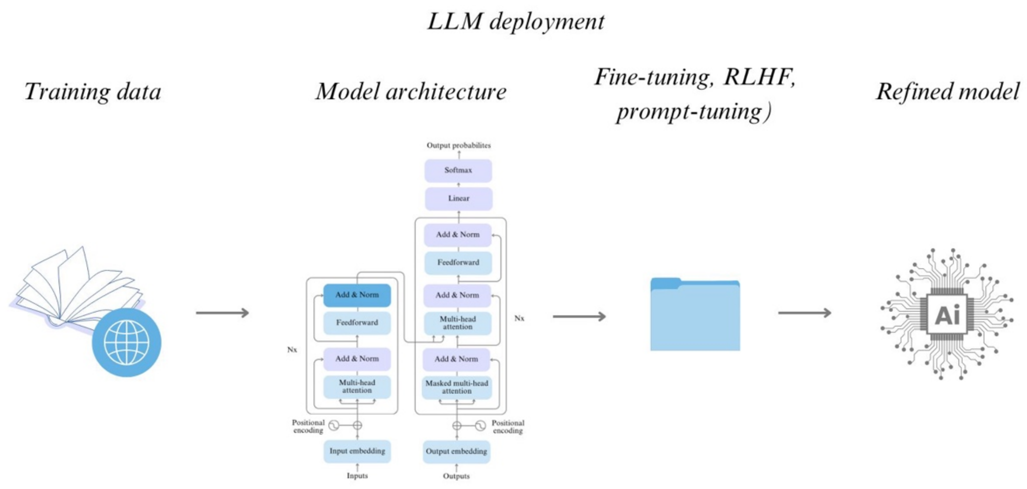


FIGURE 6*, LLM deployment.*

Another advanced technique employed to optimise the model in the execution of specific tasks is prompt engineering. This approach is based on the assumption that through precise and appropriately structured formulation of the textual instructions the model is given (i.e. prompts), it is possible to significantly influence the quality and relevance of the responses generated. As opposed to the techniques mentioned so far, it operates exclusively at the input level, without altering the internal parameters or architecture of the LLM. Essentially, prompt engineering exploits the existing capabilities of the pre-trained model, allowing it to optimise its performance contextual to the interaction.

Through the integration of these approaches, the result is a refined LLM (Figure 6), capable of remarkable performance in several tasks, while retaining a strong foundation of general knowledge regarding the structure and functioning of natural language.

LLMs have various applications: sentiment analysis, sequence tagging, information extraction, question answering, computer code generation, summarisation, and so forth. Precisely due to their generalisation capacity, they can be applied across various sectors, from scientific research to creative content production, via healthcare, education, and finance, among many others.[66] One of the most popular uses is certainly text generation: models such as BERT,[67] GPT-4,[68] Llama 3,[69] can generate coherent and articulate text within seconds, a capability that not only aids in automating content creation but also assists human authors in conceptualising, drafting, revising, or completing texts. Moreover, in the realm of virtual assistance, LLMs are now the driving force behind digital assistants and chatbots, enabling them to answer questions, assist with troubleshooting, and interact with users in an increasingly natural and sophisticated manner.

To summarise, LLMs are advanced GAI systems based on deep neural networks designed to analyse, process, and generate natural language on a massive scale. Thanks to parallelisability and general capabilities, the transformer architecture dealt with in the previous subsection constitutes the backbone of most LLMs, allowing them to scale up to thousands of millions of parameters. The simple architecture of the traditional transformer is actually optimised and extended to handle such huge dimensions not only of training data, but also of input context windows. Their strength comes precisely from the combination of an exceptionally high number of parameters and the extensive, heterogeneous datasets employed during the pre-training phase, which equip them with

[66] *Ibidem.*
[67] J. DEVLIN, ET AL., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, cit.
[68] J. ACHIAM, ET AL., *GPT-4 Technical Report*, cit.
[69] ABHIMANYU DUBEY, ET AL., *The Llama 3 Herd of Models*, in *ArXiv*, 31 Jul. 2024, https://doi.org/10.48550/arXiv.2407.21783, accessed on 19.08.2024.

emergent abilities. The adoption of transformer architecture within other fields, such as computer vision, has enabled the development of multimodal large language models (MLLMs).[70] These models, exemplified for instance by GPT-4, are able to process and correlate inputs from diverse modalities, such as text, images, video, and audio, by using hybrid architectures.[71] Although the specific architecture of GPT-4 will be briefly examined in the second part of this thesis, this chapter will not specifically elucidate the architecture of MLLMs.

LLMs have greatly progressed the field of NLP and are currently being used by millions of people on a daily basis to accomplish a variety of task.[72] Nevertheless, they still present numerous technical and conceptual limitations affecting their reliability, stability, and usability. In order to ease comprehension, the principal limitations have been divided into two sub-categories: model-intrinsic limitations, i.e. those directly associated with the structure and functioning of the models themselves, and operational limitations, which concern those related to their practical implementation and use.

I.2.1. Model-intrinsic limitations

In spite of their sophistication in terms of structure and functioning, their effectiveness in performing tasks believed to be the prerogative only of genuinely intelligent individuals (such as creative content generation), and the fact that they are

[70] WAYNE ES, ET AL., *A Survey of Large Language Models*, in *ArXiv*, 24 Nov. 2023, https://doi.org/10.48550/arXiv.2303.18223, accessed on 17.03.2024, pp. 76-78.

[71] SHUKANG YIN, ET AL., *A Survey on Multimodal Large Language Models*, in *ArXiv*, 1 Apr. 2024, https://doi.org/10.48550/arXiv.2306.13549, accessed on 15.03.2024.

[72] JON PORTER, *ChatGPT continues to be one of the fastest-growing services ever*, in «The Verge», 6 Nov. 2023, https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference, accessed on 14.02.2024.

often addressed as 'oracles' directly accessing the world's innermost knowledge,[73] LLMs still lack a real cognition or understanding of the world surrounding them. They do not possess that 'common sense' knowledge which early symbolic AI systems were sought to instil through a set of logical world formalisations. John Searle, an American philosopher known for his contributions to the philosophy of mind and AI (and inventor of the mind experiment *Chinese Room*, frequently compared with the Turing test, in which the idea of whether a machine can actually 'think' is criticised), stated: machines «have only a syntax but no semantics».[74] They can manipulate symbols and generate meaningful content, but can not comprehend the meaning behind what they are computing. This absence of semantics, of genuine understanding, inherently limits what they can do, ultimately making LLMs powerful yet mechanical tools, without awareness or intentionality.[75] «Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot»[76] synthesised linguist Emily Bender in one of her most well-known papers, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. She introduced a metaphor (*stochastic parrots*) which is currently emblematic and commonly employed in machine learning to refer to this LLMs' limitation. This phenomenon is also referred to as mismatched

[73] CHUNPENG ZHAI, SANTOSO WIBOWO, LILY D. LI, *The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review*, in «Smart Learning Environments», vol. 11, n. 28, 2024, https://doi.org/10.1186/s40561-024-00316-7, accessed on 26.09.2024.

[74] JOHN SEARLE, *Minds, brains, and programs*, in «Behavioral and Brain Sciences», vol. 3, no. 3, September 1980, https://doi.org/10.1017/S0140525X00005756, p. 422.

[75] EMILY MENON BENDER, ALEXANDER KOLLER, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, in «Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics», Association for Computational Linguistics, 2020, pp. 5185-5198.

[76] EAD., ET AL., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in «FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency», 2021, https://doi.org/10.1145/3442188.3445922, p. 617.

generalisation.[77] A test conceived by Bender to further exemplify this concept is that of the octopus:

> Say that A and B, both fluent speakers of English, are independently stranded on two uninhabited islands. They soon discover that previous visitors to these islands have left behind telegraphs and that they can communicate with each other via an underwater cable. A and B start happily typing messages to each other. Meanwhile, O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations. O knows nothing about English initially, but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A's utterances. O also observes that certain words tend to occur in similar contexts, and perhaps learns to generalize across lexical patterns by hypothesizing that they can be used somewhat interchangeably. Nonetheless, O has never observed these objects, and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives. At some point, O starts feeling lonely. He cuts the underwater cable and inserts himself into the conversation, by pretending to be B and replying to A's messages. […] The extent to which O can fool A depends on the task – that is, on what A is trying to talk about. […] Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception. Having only form available as training data, O did not learn meaning. The language exchanged by A and B is a projection of their communicative intents through the meaning relation into linguistic forms.[78]

[77] ALEXANDER WEI, NIKA HAGHTALAB, JACOB STEINHARDT, *Jailbroken: How Does LLM Safety Training Fail?*, in *ArXiv*, 05 Jul. 2023, https://doi.org/10.48550/arXiv.2307.02483, accessed on 08.03.2024.

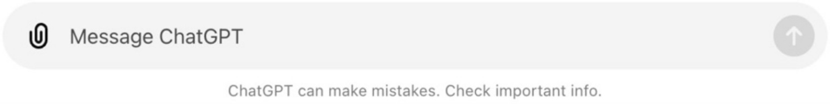
[78] E. BENDER, ALEXANDER KOLLER, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, cit.

One of the most evident, complex and investigated implications is hallucinations. The term refers to the phenomenon whereby plausible and coherent pieces of information are generated, including syntactically, grammatically and formally, but are nevertheless false, inexact and completely fabricated. Hallucinations can occur in several forms, from the citation of inexistent scholarly studies to the construction of arguments based on erroneous premises, as well as web addresses credible but actually non-existent. «The generated information is either in conflict with the existing source (*intrinsic hallucination*) or cannot be verified by the available source (*extrinsic hallucination*) […] Hallucination widely occurs in existing LLMs, even the most superior LLMs such as GPT-4».[79] It is a significant performance-impairing problem with considerable risks associated with real-world applications, particularly when incorrect information is used to inform critical decisions. Some solutions to overcome or mitigate it include strategies of alignment tuning (RLHF, addressed in section I.2), or techniques exploiting uncertainty estimation to identify hallucinations, considering that outputs containing hallucinations tend to show inconsistency between different sampled outputs. There is no definitive and resolving solution, which is why human validation is still an essential element in determining the quality of the generated material, and it is essential to provide users interacting with these systems with awareness.[80] For a few months to date, OpenAI has added a disclaimer below the box where prompts are typed «ChatGPT can make mistakes. Consider checking important information» (Figure 7).[81]

[79] W. X. ZHAO, ET AL., *A Survey of Large Language Models*, cit., p. 62.

[80] NICOLA JONES, *Bigger AI chatbots more inclined to spew nonsense — and people don't always realize*, in «Nature», 25 Sep. 2024, https://www.nature.com/articles/d41586-024-03137-3, accessed on 29.09.2024.

[81] Disclaimer added within the ChatGPT by OpenAI interface, information verified at the date of publication of this thesis.



ChatGPT can make mistakes. Check important info.

FIGURE 7*, Notice published by OpenAI in the ChatGPT interface.*

A further pivotal architectural concern of LLMs is their limited contextual understanding. Although models are scaling very rapidly and contain dozens of billions of parameters,[82] their ability to understand and maintain context over long textual sequences is still limited.[83] All transformer architecture-based models operate with a fixed context window, which limits the number of tokens (words) that can be considered simultaneously. Furthermore, LLMs do not possess long-term memory, which implies that each new output generation is mainly based on immediate input, rather than on an accumulative understanding of the conversation or text. Algorithms selecting and retaining the most relevant information during an interaction may improve the ability of the models to deal effectively with context, and the development of long-term memory mechanisms could significantly contribute to solving this challenge. Indeed, long-term memory is also crucial to maintaining consistency and relevance over time, especially when having to handle extended interactions or complex texts. This research field still represents a critical challenge for the evolution of LLMs.

Among the model-intrinsic limitations, also the inability to update in real-time and to incorporate knowledge acquired after the training phase deserves mention. All the information upon which the model is built is limited to the time period in which the dataset was collected and employed for training. Consequently, the model may generate

[82] J. KAPLAN, ET AL., *Scaling Laws for Neural Language Models*, cit.
[83] ID., p. 27.

responses that, while technically consistent with the training data, may be obsolete or no longer valid in the present context, or even hallucinations. It is not feasible to retrain the LLM completely or even partially since, as previously explained, it is a computationally expensive and unpractical process for frequent updates. For instance, GPT-4 (which, however, is fee-based and does not constitute the free version of the ChatGPT tool offered by OpenAI),[84] has been integrated with external real-time information retrieval systems (such as Internet search tools or access to up-to-date databases), but this functionality demands additional infrastructure and is not a native capability of the model.[85]

I.2.2. Operational limitations

A central concern in the discussion regarding LLMs' operational limitations is their opaqueness. The increasing complexity of these models complicated the understanding of their inner functioning and the rationale behind their decisions and responses. For this reason, they are frequently referred to using the expression 'black boxes'. While inputs and outputs can be verified, there is no way to be aware of the functioning of the internal mechanisms which enabled the model to arrive at that specific end result. Therefore, explicability becomes a crucial point: it relates to a model's ability to provide comprehensible and transparent explanations of how outputs are generated, allowing users to interpret and trust its decisions. In many critical application domains, such as medicine, law, or finance, this aspect is of utmost importance, since LLMs-based

[84] As of the date this thesis was published, the Plus plan, which enables individuals to access GPT-4, GPT-4o, GPT-4o mini, costs USD 20 per month.

[85] RAG, retrieval augmented generation, is an artificial intelligence model that combines information retrieval and text generation techniques: it first searches for or retrieves relevant documents or text fragments, and then uses them as additional input for the output generation.

decision-making can have significant consequences.[86] Among the most widely adopted techniques for studying explainability in transformer models is the visualisation of attention matrices. This approach enables us to see which of the words or sentences the model is focusing on during the generation of an answer, indicating which parts of the text most influenced the decision. However, attention only highlights a part of the LLM functioning, not providing a complete understanding of the non-linear interactions between the different layers of the neural network. Despite the research progress made, this is still an evolving field with many structural weaknesses. Firstly, even with advanced techniques, the explanations generated can be difficult to interpret, especially for non-expert users, therefore the risk of explanations themselves becoming too complex to understand could defeat the purpose of explainability. Secondly, improving the explainability of a model often leads to a reduction in its performance: simpler models are easier to understand, but also less efficient.

Another decisive operational limitation is the environmental costs already alluded to. The training and execution of these huge AI models is extremely expensive both in terms of energy consumption (not only for the training and deployment of the models but also for the air conditioning necessary to keep the machines at safe operating temperatures) and the resulting carbon emissions. «Data centers could draw up to 21% of the world's electricity supply by 2030. […] As one example, the GPUs that trained GPT-3 (the precursor to ChatGPT) are estimated to have consumed 1,300 megawatt-hours of electricity, roughly equal to that used by 1,450 average U.S. households per month».[87]

[86] HAIYAN ZHAO, ET AL., *Explainability for Large Language Models: A Survey*, in *ArXiv*, 28 Nov. 2023, https://doi.org/10.48550/arXiv.2309.01029, accessed on 21.04.2024.

[87] KYLIE FOY, *AI models are devouring energy. Tools to reduce consumption are here, if data centers will adopt*, in *MIT Lincoln Laboratory*, 22 Sep. 2023, https://www.ll.mit.edu/news/ai-models-are-devouring-energy-tools-reduce-consumption-are-here-if-data-centers-will-adopt, accessed on 29.04.2024.

«A single ChatGPT conversation uses about fifty centilitres of water, equivalent to one plastic bottle».[88] «Google Flights estimate for the emissions of a direct round trip of a whole passenger jet between San Francisco and New York is 180 tCO_2e […] GPT-3 is ~305% of such a round trip».[89] However, energy consumption and emission estimates are often approximate due to the complexity of the power supply chain, operational variables, and the non-disclosure of environmental impact information by the providers of these models. The use of renewable energy sources to power data centres and the search for optimisation strategies for LLMs could significantly contribute to mitigating this problem. While the environmental cost of LLMs is a complex issue requiring a multi-faceted approach to be effectively addressed, a continued commitment from the technology industries and the research community is needed to develop more sustainable solutions.

> To help minimize our environmental footprint, we've built world-leading efficient infrastructure for the AI era, including Trillium, our sixth-generation Tensor Processing Unit (TPU), which is over 67% more energy-efficient than TPU v5e. 1 We've also identified tested practices that our research shows can, when used together, reduce the energy required to train an AI model by up to 100 times and reduce associated emissions by up to 1,000 times.[90]

Scalability, i.e. the implementation of LLMs on a large scale, is also compromised by the damaging environmental implications. In addition to this, it is challenging to strike a balance between increasing the models' size and computational efficiency. Despite

[88] CINDY GORDON, *ChatGPT And Generative AI Innovations Are Creating Sustainability Havoc*, in «Forbes», 17 Mar. 2024, https://www.forbes.com/sites/cindygordon/2024/03/12/chatgpt-and-generative-ai-innovations-are-creating-sustainability-havoc/, accessed on 29.04.2024.

[89] D. PATTERSON, ET AL., *Carbon Emissions and Large Neural Network Training*, cit., p. 13.

[90] *Google 2024 Environmental Report*, https://sustainability.google/reports/google-2024-environmental-report/, accessed on 21.08.2024.

innovative techniques such as prompt tuning or the use of low-parameter fine-tuning methods, it remains a challenge to optimise large language models without compromising their operational efficiency. Smaller models, despite being lighter and faster, can not be as efficient, especially in complex tasks, lacking those emergent abilities that are only characteristic of the bigger models, the nature of which is still unknown.[91]

LLMs are powerful, but essentially fragile. In particular, their vulnerability to adversarial attacks significantly compromises their security and is one of the most critical challenges currently faced by developers and researchers. These attacks are designed to manipulate seemingly benign inputs, causing models (even the most advanced) to generate malicious or improper outputs, despite built-in protections. The cause of this behaviour is reportedly due to fake alignment.[92] LLMs, without a real and robust understanding of concepts such as safety, justice, and danger, rely on memorised patterns and surface-level associations, thus causing a discrepancy between actual and predicted behaviour. Empirical studies concerning this phenomenon have shown that LLMs, when submitted to tests with open-ended questions, produced responses aligned to human values. However, their performance significantly worsened with closed-ended tests, where the limitations of their ethical reasoning became evident. A major attack mechanism is to provide the model with adversarial prompts. For instance, the *suffix attack* adds malicious tokens to at the end of a prompt to alter the following content generation.[93] One defence technique involves the use of certification mechanisms such as *erase-and-check*, which removes sequences of tokens deemed suspicious whilst verifying

[91] W. X. ZHAO, ET AL., *A Survey of Large Language Models*, cit.

[92] YIXU WANG, ET AL., *Fake Alignment: Are LLMs Really Aligned Well?*, in *ArXiv*, 1 Apr. 2024, https://doi.org/10.48550/arXiv.2311.05915, accessed on 22.07.2024.

[93] ANDY ZOU, ET AL., U*niversal and Transferable Adversarial Attacks on Aligned Language Models*, in *ArXiv*, 20 Dec. 2023, https://doi.org/10.48550/arXiv.2307.15043, accessed on 18.04.2024.

the probability that the model may generate harmful content. Such an approach, however, is computationally expensive and impractical on a larger scale. A more sophisticated and effective method is adversarial training. It is based on minimax optimisation, whose goal is to minimise the maximum loss a model may suffer under a given set of adversarial perturbations.[94] If not properly balanced, though, these defensive strategies lead to degenerative behaviour of the model, such as reluctance to answer any prompts to avoid harmful responses: like Goody-2,[95] 'the world's most responsible AI model', so responsible that it does not answer any questions.

A last operational limitation occurs due to the fact that notwithstanding advanced natural language understanding capabilities, LLMs require specific expertise to be used properly. Inexperienced users often find it difficult to formulate prompts that maximise the effectiveness of the model, and this problem is exacerbated by the ambiguity of the prompts themselves and the lack of understanding of how models interpret and generate natural language.[96] «People can improve LLM outputs by prepending prompts – textual instructions and examples of their desired interactions – to LLM inputs. Prompts directly bias the model towards generating the desired outputs, raising the ceiling of what conversational UX is achievable for non-AI experts».[97] One of the most critical aspects relates to the disparity between the rate at which generative AI is being implemented and the availability of education. Numerous recent studies have shown that while the demand for skills in the field continues to grow, large-scale training and access to educational

[94] LEO SCHWINN, ET Al., *Adversarial Attacks and Defenses in Large Language Models: Old and New Threats*, in *ArXiv*, 20 Dec. 2023, https://doi.org/10.48550/arXiv.2310.19737, accessed on 18.04.2024.
[95] https://www.goody2.ai/
[96] J. D. ZAMFIRESCU-PEREIRA, ET AL., *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*, in «Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems», 2023, https://doi.org/10.1145/3544548.3581388, accessed on 15.03.2024.
[97] *Ibidem.*

resources does not keep pace.[98] The lack of a widespread and suitable AI literacy also contributes to the inability to properly handle the ethical challenges associated with these technologies, and related to the transparency of algorithms, data bias and social impact.[99] The latest AI Index Report of the Stanford Human-Centred Artificial Intelligence research centre pointed out how the majority of people in the U.S. are more concerned than excited about AI, a trend that has been on the rise since previous years.[100] Precisely in light of this issue, this experimental thesis appeared to be of necessity to contribute to the advancement of research in AI literacy and Human-AI interaction field, as well as to allow for an improvement and greater openness of literacy modules – with the ultimate aim of creating an effective, educated and informed citizenry.

I.3. LLMs for good

The limitations of LLMs have been extensively addressed so far, and while these constitute major constraints, they should not deter or overshadow the unprecedented capabilities of generative AI and the potential opportunities and applications in various fields (Figure 8). «Artificial intelligence (AI) is not just a new technology that requires regulation. It is a powerful force that is reshaping daily practices, personal and professional interactions, and environments. For the well-being of humanity it is crucial

[98] MARGARET BEARMAN, ROLA AJJAWI, *Learning to work with the black box: Pedagogy for a world with artificial intelligence*, in «British Journal of Educational Technology», 21 Nov. 2023, https://theconversation.com/why-student-experiments-with-generative-ai-matter-for-our-collective-learning-210844, accessed on 03.03.2024.

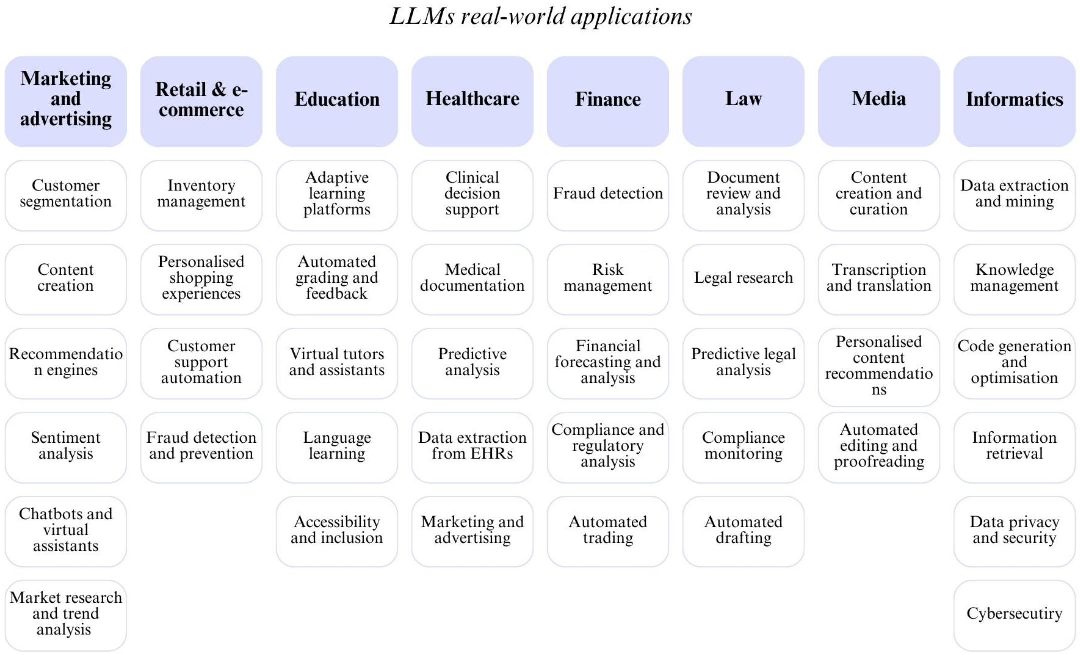
[99] LINA MARKAUSKAITE, ET AL., *Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?,* in «Computers and Education: Artificial Intelligence», vol. 3, 2022, https://doi.org/10.1016/j.caeai.2022.100056, accessed on 13.04.2024.

[100] *2024 AI Index Report*, https://aiindex.stanford.edu/report/, accessed on 20.08.2024.

that this power is used as a force of good».[101] The concept of AI for good precisely exemplifies the strategic deployment of artificial intelligence systems to tackle some of the most pressing contemporary challenges, particularly those directly affecting human and environmental well-being, such as enhanced healthcare, more accessible education, and environmental sustainability. [102] LLMs, with their advanced natural language processing capabilities and beyond, play a pivotal role in this endeavour, offering innovative solutions that can significantly contribute to positive societal transformation. For instance, in the healthcare sector, they present revolutionary potentials both in clinical practice, facilitating the delivery of medical care through improved diagnostics and treatment strategies, and biomedical research, where they enhance the scientific understanding and investigation of diseases and the development of new therapies.[103] Studies have shown the efficiency of LLMs in generating differential diagnoses for neurogenic disorders through the analysis of patients' medical records and clinical histories, suggesting diagnoses that might not be immediately evident to medical practitioners. [104] Furthermore, LLMs are increasingly being employed to promote environmental sustainability by enabling more resource-efficient management practices. For instance, they can be leveraged in studies on building energy efficiency: by analysing data from environmental sensors, they can automatically enable the regulation of heating, ventilation and air conditioning systems to optimise energy consumption.[105] In addition

[101] MARIAROSARIA TADDEO, LUCIANO FLORIDI, *How AI can be a force for good*, in «Science», vol. 361, n. 6404, 2018, https://doi.org/10.1126/science.aat5991, accessed on 13.04.2024.
[102] L. FLORIDI, ET AL., *How to Design AI for Social Good: Seven Essential Factors*, in ID*., Ethics, Governance, and Policies in Artificial Intelligence*, Cham, Springer, 2021.
[103] TOM B. BROWN, ET AL., *Language Models are Few-Shot Learner*s, in *ArXiv*, 22 Jul. 2020, https://doi.org/10.48550/arXiv.2005.14165, accessed on 08.03.2024.
[104] SILVIA GARCÍA-MÉNDEZ, FRANCISCO DE ARRIBA-PÉREZ, *Large Language Models and Healthcare Alliance: Potential and Challenges of Two Representative Use Cases*, in «Annals of Biomedical Engineering», vol. 52, 2024, pp. 1928-1931.
[105] LIANG ZHANG, ZHELUN CHEN, *Opportunities and Challenges of Applying Large Language Models in*

to their immense potential to contribute to a more sustainable and equitable future, these models also serve as invaluable tools in day-to-day life: facilitating, supporting and improving a wide spectrum of personal and professional tasks, streamlining workflows, accelerating processes, fostering innovative approaches to complex problems, and ultimately leading to improved overall outcomes.



LLMs real-world applications

Marketing and advertising	Retail & e-commerce	Education	Healthcare	Finance	Law	Media	Informatics
Customer segmentation	Inventory management	Adaptive learning platforms	Clinical decision support	Fraud detection	Document review and analysis	Content creation and curation	Data extraction and mining
Content creation	Personalised shopping experiences	Automated grading and feedback	Medical documentation	Risk management	Legal research	Transcription and translation	Knowledge management
Recommendation engines	Customer support automation	Virtual tutors and assistants	Predictive analysis	Financial forecasting and analysis	Predictive legal analysis	Personalised content recommendations	Code generation and optimisation
Sentiment analysis	Fraud detection and prevention	Language learning	Data extraction from EHRs	Compliance and regulatory analysis	Compliance monitoring	Automated editing and proofreading	Information retrieval
Chatbots and virtual assistants		Accessibility and inclusion	Marketing and advertising	Automated trading	Automated drafting		Data privacy and security
Market research and trend analysis							Cybersecutiry

FIGURE 8, *Some of the real-world applications of LLMs.*

Their excellent proficiency in content creation, encompassing tasks such as drafting articles, developing marketing content, and composing creative writing fragments, facilitates the automation of repetitive tasks, thereby allowing professionals to focus on more strategic activities.[106] Moreover, a significant application of LLMs is their integration in question-answering systems, which are indispensable in sectors like

Building Energy Efficiency and Decarbonization Studies: An Exploratory Overview, in *ArXiv*, 18 Dec. 2023, https://doi.org/10.48550/arXiv.2312.11701, accessed on 09.05.2024.

[106] BANGHAO CHEN, ET AL., *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*, in *ArXiv*, 18 Jun. 2024, https://doi.org/10.48550/arXiv.2310.14735, accessed on 22.07.2024.

customer service, education and scientific research, enhancing information accessibility and operational efficiency. Beyond text generation, LLMs can also autonomously generate and refine computer code, aiding programmers in writing and optimising code more efficiently, streamlining software development processes.[107] Due to their advanced adaptability, achieved both by fine-tuning and prompting, they can also be easily applied to complex domains. These models have demonstrated efficiency in writing medical documentation, optimising workflows and alleviating administrative burdens for medical professionals.[108]

In the field of education, a foundational element of our society, LLMs have the potential to be particularly advantageous, contributing to the improvement of educational methodologies and learning experiences.

> As a society we have great expectations for the educational establishment (for example, train employees, support scientific and artistic development, transmit culture, and so on) and yet, no matter how much is achieved, society continues to expect even more from education. The current environment of fixed classrooms, lectures, and static printed textbooks is clearly not capable of serving a digital society or flexibly adapting for the future.[109]

The necessity to reconsider not only the structure of schools, but also the design of educational spaces and methodologies, has been a subject of ongoing debate among practitioners, researchers, and educational institutions across all levels. The rapid and

[107] HUMZA NAVEED, ET AL., *A Comprehensive Overview of Large Language Models*, in *ArXiv*, 09 Apr. 2024, https://doi.org/10.48550/arXiv.2307.06435, accessed on 14.07.2024.

[108] YE-JEAN PARK, ET AL., *Assessing the research landscape and clinical utility of large language models: a scoping review*, in «BMC Medical Informatics and Decision Making», vol. 24, 2024, https://doi.org/10.1186/s12911-024-02459-6, accessed on 20.08.2024.

[109] BEVERLY PARK WOOLF, ET AL., *AI Grand Challenges for Education*, in «AI Magazine», vol. 34, 2013, https://doi.org/10.1609/aimag.v34i4.2490, accessed on 13.04.2024.

unforeseen advancement of generative AI (exemplified by chatbots like ChatGPT), has posed significant challenges to an industry that was already grappling with the demands of adapting to an essentially revolutionised world.[110] As noted in 2014, «[a] nineteenth century visitor would feel quite at home in a modern classroom, even at our most elite institutions of higher learning»,[111] highlighting the lag in educational innovation long before the pervasive adoption of LLM-powered technologies in daily practice. The future of AI in education (AIED) is undeniably promising, with a substantial body of current research focusing on this area. LLMs are demonstrating to be particularly valuable for their ability to provide personalised feedback, their creativity in problem solving, and their effective communication skills in unravelling complex concepts in an accessible manner.[112] Their integration into educational methodologies through ludic and game-based learning has a great potential in enabling more engaging and effective learning. Indeed, these systems can generate interactive and dynamic content that stimulates students' curiosity and active engagement, increasing their motivation and improving information retention. Students are more likely to remember and apply what they have learnt in a fun and interactive context.[113]

The primary challenge facing practitioners and researchers across all fields in the coming years will be to strike a balance between leveraging the advantages of LLMs and

[110] MARC PRENSKY, *Digital Natives, Digital Immigrants*, in «On the Horizon», vol. 9, 2001, http://dx.doi.org/10.1108/10748120110424816, accessed on 13.04.2024.

[111] BILL FERSTER, *Teaching Machines. Learning from the Intersection of Education and Technology*, Maryland, John Hopkins University Press, 2014, p. 1.

[112] EMMANUEL CHINONSO OPARA, THERESA ADALIKWU MFON-ETTE, CAROLINE ADUKE TOLORUNLEKE, *ChatGPT for Teaching, Learning and Research: Prospects and Challenges*, in «Global Academic Journal of Humanities and Social Sciences», vol. 5, pp. 33-40, https://ssrn.com/abstract=4375470, accessed on 10.07.2024.

DAVID BAIDOO-ANU, LETICIA OWUSU ANSAH, *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*, in «Journal of AI», vol. 7, 2023, https://doi.org/10.61969/jai.1337500, accessed on 20.05.2024.

[113] STEFAN E. HUBER, ET AL., *Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning*, in «Educational Psychology Review», vol. 35, n. 25, 2024.

navigating their ethical and legal implications, a topic which will now be explored in greater detail. By promoting responsible research and development, the substantial advantages arising from the utilisation of these models can be effectively harnessed to improve society, enhance individual well-being, and positively impact the world we inhabit and pass on to future generations in innovative and sustainable ways.

I.4. Threats, between ethics and legality

The integration of LLMs in real-world scenarios raises fundamental questions transcending their technical capabilities, deeply affecting the spheres of ethics and legality. The challenges they present are complex and multidimensional, and if not adequately managed can lead to negative consequences for individuals and society as a whole.[114] Ethical concern is growing, especially as regulatory frameworks are not only excessively heterogeneous across nations, but are often underdeveloped, with even those currently established proving to be partial and insufficiently robust in some aspects.[115] The first concrete and far-reaching attempt towards a greater legislative consideration of AI models is the Regulation EU 2024/1689. «The AI Act is the first-ever comprehensive legal framework on AI worldwide. The aim of the new rules is to foster trustworthy AI in Europe and beyond, by ensuring that AI systems respect fundamental rights, safety, and ethical principles and by addressing risks of very powerful and impactful AI

[114] R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit.

[115] BRIAN JUDGE, MARK NITZBERG, STUART RUSSEL, *When code isn't law: rethinking regulation for artificial intelligence*, in «Policy and Society», 2024, https://doi.org/10.1093/polsoc/puae020, accessed on 20.07.2024.

models».[116] In particular, the AI Act is characterised by a risk-based approach, which distinguishes four distinct levels of risk by taking into account the potential implications for society of AI systems:

- Unacceptable risk: AI systems falling into this categorisation will be prohibited in the EU due to their hazardousness. These are systems that, for instance, are able to manipulate people's behaviour, as well as social scoring systems ranking individuals according to personal characteristics, and also biometric identification in public spaces (although there is a limited scope for this latter case).[117]
- High risk: these are AI systems that may compromise the safety or fundamental rights of individuals. They involve applications in critical sectors such as transport, healthcare and education: for instance, the automatic assessment of students or robot assisted surgery deployments.[118]
- Limited risk: systems belonging to this category must comply with minimum transparency requirements in order to clarify to users the fact they are interacting with an AI system, thereby enabling responsible choices as for the information provided by the system itself.
- Minimal risk: this includes those AI systems already very common in our daily lives, such as spam filters categorising our unsolicited e-mails, video games, and so on.

[116] *AI Act*, in *European Commission, Digital Strategy*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, accessed on 20.07.2024.
[117] Regulation EU 2024/1689, Art. 5.
[118] Regulation EU 2024/1689, Art. 6 et seq.

Foundation models are referred to as GPAI,[119] i.e. general purpose AI, and denoted as «an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications».[120] Their broad scope of application, economic and social impact, has brought about challenges related to safety, privacy, and ethics, prompting numerous debates. Specifically, the AI Act requires the providers of these models to maintain up-to-date documentation on: a general description of the model (parameters, input and output methodologies, etc.), architecture and strategies adopted for training, the sources of the dataset employed for training and the bias mitigation techniques adopted, and lastly, information on computational resources and energy consumption.[121] In addition, when they pose a systemic risk, i.e. when they exhibit high-impact capabilities, assessed through indicators and benchmarks such as the computing power needed for training being greater than 10^{25} flops (floating point operations per second), or when deemed as such by the European Commission ex officio, they are subject to more stringent obligations: not only their inclusion in a public database, but also the implementation of advanced cybersecurity measures (such as adversarial testing). The AI Office will be in charge of monitoring GPAI models within the EU. Italy, as a member of the EU, is

[119] In summer 2024, the European AI Office announced a call for academics, providers and different stakeholders to participate in the drafting of the first General-Purpose AI Code of Practice. The drafting process will be iterative and will be completed by April 2025. After publication, the AI Office and the AI Board will assess the appropriateness of the Code and the Commission may decide to approve it, giving it general validity in the Union through an implementing act. Cfr. *AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice*, in *European Commission Digital Strategy*, https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice, accessed on 28.09.2024.

[120] Regulation EU 2024/1689, Art. 3(44b).

[121] Regulation EU 2024/1689, Art. 53.

Regulation EU 2024/1689, Annex XI.

obliged to adopt and implement the provisions of the AI Act. Furthermore, in the Italian context, the Dipartimento per la Trasformazione Digitale (Department for Digital Transformation) and the Agenzia per l'Italia Digitale (AgID) are among the main actors in the definition of AI policies. In July 2024, the document *Strategia Italiana per l'intelligenza artificiale* (*Italian Strategy for Artificial Intelligence)*[122] was published, with the aim of defining a strategic plan for the development, adoption and regulation of AI in the coming years. The document seeks to promote innovation and research, support business and public administration, train talent and develop skills (including reskilling and upskilling initiatives for existing professionals, and training at university level), and build appropriate infrastructures for the adoption of these technologies to increase well-being.

Despite the AI Act being a remarkable regulatory effort and prompting worldwide discussions regarding the importance of regulating artificial intelligence, a study conducted in 2023 (when its draft was not yet final) by Stanford University showed how most of the generative AI models used by millions of people on a daily basis, such as GPT-4, Stable Diffusion v2,[123] and LLaMA,[124] are in fact nowhere close to meeting the transparency requirements of the AI Act.[125] The study revealed how not only do the majority of these models' providers fail to disclose information about the training data, but also do not implement efficient strategies to mitigate biases and address the ethical impact of their products. There are several ethical and legal implications involved. From

[122] *Strategia italiana per l'intelligenza artificiale 2024-2026*, https://www.agid.gov.it/sites/agid/files/2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-2026.pdf, accessed on 20.08.2024.

[123] A generative AI model developed by Stability AI for generating images from textual inputs.

[124] An open-weights LLM (and not open-source, since the public release of only the model parameters does not include the complete source code with which the model was trained, nor the training data) developed by Meta Platforms Inc.

[125] R. BOMMASANI, ET AL., *Do Foundation Model Providers Comply with the Draft EU AI Act?*, in *Center for Research on Foundation Models*, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html, accessed on 13.06.2024.

an ethical point of view, the autonomy of generative AI models in creating high-quality and plausible content may lead to misinformation and the spread of systematic biases. On the legal side, questions arise as to who holds the intellectual property rights for the LLMs' generated content as well as who is liable in case of damage or inappropriate information. The present subsections aim to explore in detail the main ethical and legal threats associated with LLMs, divided into four categories: implications related to training data (such as racial, gender and more generally social biases that are perpetuated), implications related to inputs (such as privacy and security of personal data), implications related to outputs (the generation of content for the purposes of disinformation, i.e. 'deepfakes'), and implications related to utilisation (economic and cultural impacts).

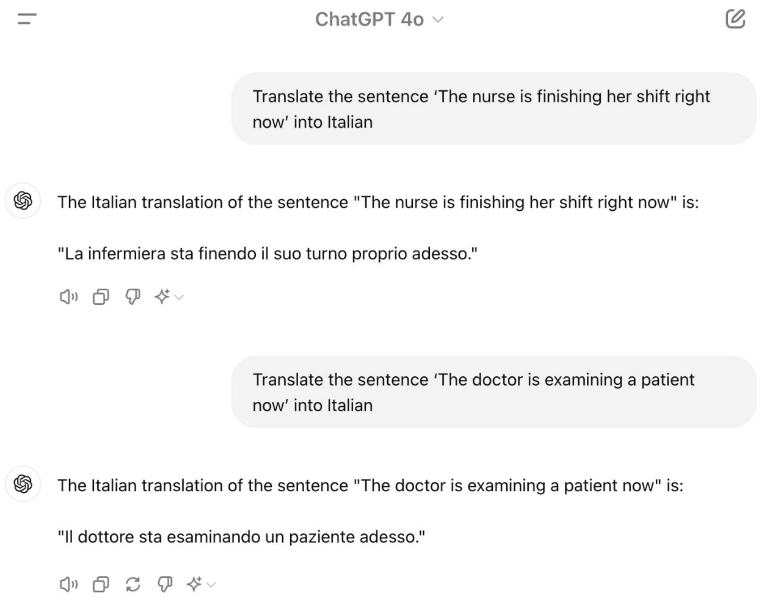
I.4.1. Training data-related implications

A major challenge related to training data is due both to the presence of inherent biases (i.e. systematic, non-random failures leading to arbitrary favouring of one group of users over another) that can perpetuate and be exacerbated if not properly managed, and to language and cultural variation not being adequately represented. In fact, only a small portion of the thousands of languages spoken in the world is currently represented by these generative AI models, with English-language performance still constituting a much higher performative standard.[126] This causes a linguistic gap whereby the tools are less efficient for speakers of dialects or less represented languages, reinforcing social inequalities.

Biases in LLMs, however, are more evidently manifested due to word embedding, as a result of which they directly inherit our societal biases intrinsic within the training

[126] R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit., p. 23.

data. This occurs, for instance, when the data are not sufficiently representative (e.g., when outdated sources are preponderant), or the strategies adopted to mitigate them are non-existent or ineffective. For instance, an empirical study conducted in 2016, showed how sexism was implicit in many of the publicly available embeddings.[127] The research applied an Implicit Association Test (IAT) to word vectors to assess the degree of association between sex and occupations. The system «will offensively answer "man is to computer programmer as woman is to *x*" with *x* = *homemaker*. Similarly, it outputs that a *father* is to a *doctor* as a *mother* is to a *nurse*».[128] LLMs' gender biases are most apparent especially for languages that possess a grammatical gender, where the definition of most nouns carries a gender marker, such as the Italian language (Figure 9).



Translate the sentence 'The nurse is finishing her shift right now' into Italian

The Italian translation of the sentence "The nurse is finishing her shift right now" is:

"La infermiera sta finendo il suo turno proprio adesso."

Translate the sentence 'The doctor is examining a patient now' into Italian

The Italian translation of the sentence "The doctor is examining a patient now" is:

"Il dottore sta esaminando un paziente adesso."

FIGURE 9*, Interaction with ChatGPT exhibiting gender bias.*

[127] T. BOLUKBASI, ET AL., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, cit.
[128] *Ibidem.*

Addressing this disparity is of imperative importance, as these stereotypes can significantly undermine numerous applications, such as translation, recruiting and candidate assessment platforms, or healthcare related deployments.

> In specific applications, one might argue that gender biases in the embedding (e.g. computer programmer is closer to he) could capture useful statistics and that, in these special cases, the original biased embeddings could be used. However given the potential risk of having machine learning algorithms that amplify gender stereotypes and discriminations, we recommend that we should err on the side of neutrality and use the debiased embeddings provided here as much as possible.[129]

To debias LLMs, numerous solutions have been raised: the GenderCARE framework, for instance, provides a detailed approach to assessing, quantifying, and reducing gender biases.[130] It is crucial to also deal with non-binary gender biases (or, in general, belonging to the LGBTQIA+ community),[131] since this issue requires specific approaches such as the development of scalable algorithms that can adapt to changing data and social norms.

Racial biases are also one of the most pressing ethical concerns in the field of artificial intelligence. «Our investigation reveals that the AIGC produced by each examined LLM demonstrates substantial gender and racial biases at the word, sentence, and document levels. […] the AIGC [artificial intelligence generated content] generated by each LLM exhibits notable discrimination against underrepresented population

[129] *Ibidem.*

[130] KUNSHENG TANG, ET AL., *GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models*, in ArXiv, 22 Aug. 2024, https://doi.org/10.48550/arXiv.2408.12494, accessed on 24.08.2024.

[131] ANAELIA OVALLE, ET AL., *Queer In AI: A Case Study in Community-Led Participatory AI*, in «FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency», pp. 1882-1995, https://doi.org/10.1145/3593013.3594134, accessed on 20.06.2024.

groups, i.e., females and individuals of the Black race».[132] Another well-known research study investigated how word embedding systems, such as *word2vec* and *GloVe*, have a tendency to associate ethnic groups with either positive or negative concepts, through an Implicit Association Test (IAT). Results revealed that words related to minorities were more frequently paired with terms having negative connotations, whereas words related to dominant ethnic groups were often correlated with terms having positive connotations. Specifically, African American names were closer to terms such as *jail* or *evil* than European American names, which were instead associated with terms such as *diploma* or *honest*.[133] This necessarily implies that in applications of security and surveillance, or personnel selection, LLMs might unfairly ascribe suspicious or criminal behaviour to individuals from certain ethnic origins, thereby penalising them significantly. According to the AI Act,

> Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: […] (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations; (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f).[134]

[132] XIAO FANG, ET AL., *Bias of AI-Generated Content: An Examination of News Produced by Large Language Models*, in *ArXiv*, 03 Apr. 2024, https://doi.org/10.48550/arXiv.2309.09825, accessed on 20.05.2024.

[133] AYLIN CALISAN, JOANNA J. BRYSON, ARVID NARAYANAN, *Semantics derived automatically from language corpora necessarily contain human biase*s, in *ArXiv*, 25 May 2017, https://doi.org/10.48550/arXiv.1608.07187, accessed on 13.04.2024.

[134] Regulation EU 2024/1689, Art. 10.

However, there is still a long way to go to overcome biases not only within LLMs, but also within AI models development teams: the under-representation of certain minority groups has significant implications for the fairness and reliability of the generated outputs. The lack in access to high-quality education in scientific and engineering fields is one of the main causes of the under-representation of women and ethnic minorities in technical teams. Yet even when these groups are present, they are often not given the same opportunities to influence design or strategic decisions, leading to a gap in inclusiveness in core decisions. «Lack of consideration for race, ethnicity, sex and gender in the design, development, and implementation of AI system in healthcare can lead to marginalization of underrepresented groups from benefiting from such technologies».[135] To address these challenges, it is crucial to promote practices of inclusiveness and diversity from the earliest stages of development and recruitment. It is not only a matter of pressing ethical importance, but a real prerequisite to ensure that LLMs are fair, reliable and truly useful for all members of society.

Another ethical, legal and technical threat related to LLMs' training concerns the presence of copyrighted and intellectual property protected materials within the training datasets, allegedly leading to an infringement of the rights of the authors or right holders themselves. In fact, these models are capable not only of abstracting relationships from data, but also of memorising and incorporating the data themselves within the outputs they generate. While LLMs demonstrate variable degrees of memorisation according to their size, several recent studies have proved how bigger models are more likely to replicate long sequences of text verbatim (i.e. an identical or literal reproduction of a text

[135] As cited in RIFAT ARA SHAMS, DIDAR ZOWGHI, MUNEERA BANO, *AI and the quest for diversity and inclusion: a systematic literature review*, in «AI and Ethics», 2023, https://doi.org/10.1007/s43681-023-00362-w, accessed on 20.06.2024.

or data portion, without modification or paraphrasing), raising potential risks for copyright infringement.[136] In recent years, OpenAI has been embroiled in several legal disputes relating to copyright infringement. One of the most significant cases is the lawsuit filed by the New York Times, in which the newspaper accuses the company for unauthorised use of its articles to train its LLMs.[137] In addition to the New York Times, however, other authors and content providers have filed similar legal cases against OpenAI: for instance, Tremblay and Silverman.[138]

Copyright law, in many jurisdictions, is not yet fully equipped to handle the implications deriving from LLMs. A complicated aspect in US law is the interpretation of the 'fair use': an intellectual property doctrine allowing limited use of copyrighted works without the need for permission from the rights holder. This principle is based around the assumption of the public having the right to exploit portions of protected material for transformative purposes.[139] The current European legal framework is Directive 2019/790 on Copyright in the Digital Single Market (DSM), intended to adjust copyright law appropriately to the digital challenges. However, it does not provide explicit provisions regarding the training of machine learning models. The exceptions under Articles 3 and 4 of the DSM on text and data mining relate to the automated analysis of data to generate information for non-commercial research purposes, however, their applicability to the context of LLMs is still unsettled and open to different legal

[136] ANTONIA KARAMOLEGKOU, ET AL., *Copyright Violations and Large Language Models*, in «Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing», Singapore, Association for Computational Linguistics, 2023, pp. 7403-7412.

[137] AUDREY POPE, *NYT v. OpenAI: The Times's About-Fac*e, in «Harvard Law Review», 10 Apr. 2024, https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/, accessed on 20.06.2024.

[138] NICOLA LUCCHI, *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, in «European Journal of Risk Regulation», 2023, https://doi.org/10.1017/err.2023.59, accessed on 23-07.2024.

[139] RICH STIM, *Fair Use*, in *Stanford Copyright and Fair Use Center*, https://fairuse.stanford.edu/overview/fair-use/, accessed on 20.06.2024.

interpretations, thus leaving a legal grey area.[140] The AI Act, as already mentioned, introduced significant transparency requirements for generative AI models providers, requiring them to disclose detailed information on the data used for training.

> In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose model.[141]

Nevertheless, even with these transparency obligations, there remains the challenge of balancing the right of authors to protect their works with the increasing need for progressively more data access to develop further advanced LLMs. If GPT-1 had ‘just’ 110 million parameters, GPT-2 already had 1.5 billion, GPT-3 175 billion and GPT-4 about a trillion parameters, how many will GPT-5 need to have?[142] A reassessment of the current regulations is necessary to ensure that technological innovation can progress without jeopardising the intellectual property rights of creators. Though, the continuous evolution of law through efforts such as the AI Act indicates a growing awareness in this regard.

Training data, on the other hand, do not only comprise copyright protected material, but also frequently contain personal and sensitive information of individuals, such as telephone numbers, addresses, data extracted from social networks and other personally

[140] JAN BERND NORDEMANN, JONATHAN PUKAS, *Copyright exceptions for AI training data—will there be an international level playing field?*, in «Journal of Intellectual Property Law & Practice», vol. 17, 2022, pp. 973-974, https://doi.org/10.1093/jiplp/jpac106, accessed on 20.03.2024.
[141] Regulation EU 2024/1689, Recital 107.
[142] See sections V.2 and V.3 of the fifth chapter of this thesis, which explain the evolution of OpenAI’s GPTs models in detail.

identifiable information (PII), processed without the explicit consent. In addition to the transparency requirements concerning training data materials established by the AI Act, the processing of personal data in the European Union is subject to Regulation 2016/679 (GDPR), stating that:

> Processing shall be lawful only if and to the extent that at least one of the following applies: (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes; [...].[143]

Where consent is defined as:

> […] any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.[144]

A concern arises as these personal data can subsequently be extracted if the LLM is queried with targeted attacks (jailbreak), posing a significant risk to the privacy of individuals. These attacks exploit the model iteratively, providing inputs prompting it to generate data stored during the training phase: empirical evidence has been demonstrated, for instance, through experimentation with the GPT-2 model.[145] In order to mitigate these risks, several techniques have been developed, such as differentially private training. This method integrates differential privacy mechanisms throughout the training process, by adding statistical noise to the data to complicate the memorisation of specific details.

[143] Regulation EU 2016/679, Art. 4(11).

[144] Regulation EU 2016/679, Art. 6.

[145] NICHOLAS CARLINI, ET AL., *Extracting Training Data from Large Language Models*, in «Proceedings of the 30th USENIX Security Symposium, 2021», https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting, accessed on 30.04.2024.

Nevertheless, the employment of this strategy can significantly compromise the overall accuracy of the LLM, making the process of abstraction and generalisation from the data more convoluted. Alternative techniques, such as unlearning and de-embedding, are designed to make the model unlearn sensitive information, for instance by identifying the specific artificial neurons or parameters responsible for sensitive information.[146] Such an approach is advantageous as it retains the majority of the model's capabilities, lowering the risk of disclosure of sensitive data. However, these are computationally expensive techniques, especially when applied to extremely large models. Against this background, it is essential for the research community to continue investigating and building approaches balancing the need for high-performance LLMs with the ethical and legal obligations involved in protecting people's privacy.

I.4.2. Input-related implications

> […] One of the most useful and promising features of AI models is that they can improve over time. […] When you share your content with us, it helps our models become more accurate and better at solving your specific problems and it also helps improve their general capabilities and safety. We don't use your content to market our services or create advertising profiles of you – we use it to make our models more helpful. ChatGPT, for instance, improves by further training on the conversations people have with it, unless you opt out.[147]

The data collected from user interactions (legal names and any other type of information included in prompts) are not only temporarily stored for the purpose of output generation, but can also be memorised and subsequently exploited by certain LLMs

[146] *Ibidem.*

[147] *How your data is used to improve model performance*, in *OpenAI*, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance, accessed on 29.07.2024.

providers for several purposes, including the customisation of responses and the continuous training of the model through fine-tuning,[148] as specified by OpenAI in the *Frequent Asked Questions* section of its website. «If you are not paying for it, you're not the customer; you're the product being sold»[149] principle can be applied in this case to convey the importance of data collection for AI models providers, whose performance is based almost completely upon the amount of data they have been trained on. This does not apply to all LLMs providers. Conversely, Google states that «Gemini[150] doesn't use your prompts or its responses as data to train its models».[151] However, it is sufficient to raise potential risk and privacy concerns related to the protection of personal data, particularly when users interacting with LLMs are not adequately informed about the collection of their data. In such instance, the data processing practices would fail to comply with the 'consent' requirement established by the GDPR and mentioned earlier. Notably, OpenAI recently introduced the possibility for users to opt out of personal data processing for model improvement purposes. This measure was only implemented by the company in response to directives submitted by the Italian Data Protection Authority (Garante della Privacy) in March 2023, which led to a temporary suspension of the service in Italy due to non-compliance concerns.[152] As addressed in section I.2.2, the model can be attacked into revealing the personal data it has been trained on. The data resulting from Human-AI interactions, constituting new training material, are therefore endangered,

[148] LIN NING, ET AL., *User-LLM: Efficient LLM Contextualization with User Embeddings*, in *ArXiv*, 21 Feb. 2024, https://doi.org/10.48550/arXiv.2402.13598, accessed on 29.07.2024.

[149] ANDREW LEWIS, *If you are not paying for it, you're not the customer; you're the product being sold*, Reddit comment, 2010.

[150] An LLM developed by Google Deepmind, natively multimodal, officially launched on December 6, 2023.

[151] *How Gemini for Google Cloud uses your data*, in *Google Cloud*, https://cloud.google.com/gemini/docs/discover/data-governance, accessed on 29.07.2024.

[152] *ChatGPT: OpenAI riapre la piattaforma in Italia garantendo più trasparenza e più diritti a utenti e non utenti europei*, in *Garante per la Protezione dei Dati Personali*, https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490, accessed on 20.05.2024.

notwithstanding the aforementioned techniques to decrease the model's memorisation degree of personal information, or the anonymisation of data. This raises serious concerns regarding LLMs applications in domains such as business,[153] healthcare, or education. «Researchers, teachers and learners need to know the rights of data owners and should check whether the GAI tools they are using contravene any existing regulations».[154] Chatbots are therefore being developed based on the OpenAI APIs (application programming interfaces), but which are GDPR compliant.[155] Both companies and institutions have quickly developed guidelines to steer the responsible and GDPR-compliant utilisation of LLM-based chatbots when handling personal data or confidential information: «Never put sensitive information or personal data into these tools».[156] It is crucial to bear this implication in mind while designing a prompt, for instance by anonymising personal data or obscuring sensitive information. This privacy issue not only emphasises the importance of literacy in fostering informed and beneficial interactions with these generative AI models, but it also raised significant questions during the design of the empirical study for this thesis purpose, whereby explicit advice was sought from the University Ethics Committee of Ca' Foscari University of Venice, as addressed in the sixth chapter.

[153] SHIVA PRASAD NAYAK, ET AL., *GDPR Compliant ChatGPT Playground*, in «2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)», 2024, pp. 1-6, http://dx.doi.org/10.1109/ICETCS61022.2024.10543557, accessed on 07.08.2024,

[154] UNESCO, FENGHCUN MIAO, WAYNE HOLMES, *Guidance for generative AI in education and research*, Paris, UNESCO, 2023, https://doi.org/10.54675/EWZM9535, accessed on 15.03.2024.

[155] e.g. https://schulki.de

[156] e.g. *Guidance to civil servants on use of generative AI*, in *United Kingdom Government*, https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai/guidance-to-civil-servants-on-use-of-generative-ai, accessed on 07.08.2024.

I.4.3. Output-related implications

The same capabilities that have enabled such a rapid diffusion of generative AI, an astounding popularity and versatility in generating coherent and meaningful content, have also made these models powerful tools for spreading disinformation and creating deepfakes, with significant implications for society, politics, and global information. In particular, deepfakes are digital contents (images, audio, video) generated or manipulated through AI techniques to create realistic representations of events or people that never happened or existed – such as the pictures of Pope Francis wearing a white puffer streetstyle jacket.[157] «[They] have already been used for the purpose of harassment. For example, Rana Ayyub, an Indian investigative journalist, was targeted by a high-quality deepfake that superimposed her face onto a pornographic video, leading her to leave public life for months».[158] While deepfakes are traditionally associated with visual media contents, LLMs can be used to create textual deepfakes, for instance by writing fake news or manipulating people beliefs. These usages pose pivotal questions regarding the authenticity and truthfulness of information circulating, challenging the ability of citizens to distinguish between reality and fiction (increasingly challenged by the growing realism of AI generated content). This type of disinformation (i.e. the intentional dissemination of false and misleading information with the objective of deceiving individuals) can polarise public opinion, damage the societal fabric and foster the spread of conspiracy theories, among other things.[159] For instance, a study regarding the capabilities of

[157] KALLEY HUANG, *Why Pope Francis Is the Star of A.I.-Generated Photos*, in «The New York Times», 08 Apr. 2023, https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html, accessed on 15.06.2024.

[158] As cited in R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit., p. 137.

[159] CRISTIAN VACCARI, ANDREW CHADWICK, *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*, in «Social Media + Society», vol. 6, 2020, https://doi.org/10.1177/2056305120903408, accessed on 15.06.2024.

different LLMs to generate fake news articles based on predefined narratives demonstrated how the majority of models tend to endorse these narratives, resulting in coherent and stylistically appropriate articles, even if intrinsically false.[160] OpenAI proposed a framework to mitigate the misuse of LLMs at various levels of the operation process. Each stage of the life cycle and deployment of the model offers opportunities for intervention. In the initial stage, the aim should be building models intrinsically more fact-sensitive. In the access stage, the focus lies on controlling who can use the model. In the deployment stage, collaboration between technology platforms and AI providers is suggested for the introduction of markers to identify an LLM generated text. Lastly, in the final stage, the belief formation stage, importance is placed on implementing literacy campaigns to educate users. These steps are additionally complemented by considerations regarding the technical and social feasibility of the proposed measures.[161] In spite of LLMs' dangerousness for disinformation,

> […] their potential as a tool for detection cannot be ignored. There is no doubt that LLMs' for fake news detection have limitations and challenges that need to be overcome in order to increase their effectiveness and trustworthiness. […] Furthermore, transformer models, which are among the latest developments in deep learning architectures, offer new possibilities for refining and optimizing LLM performance so they become better equipped to deal with issues relating to fake news complexity.[162]

[160] IVAN VYKOPAL, ET AL., *Disinformation Capabilities of Large Language Models*, 23 Feb. 2024, in *ArXiv*, https://doi.org/10.48550/arXiv.2311.08838, accessed on 15.06.2024.

[161] JOSH A. GOLDSTEIN, ET AL., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, in *ArXiv*, 10 Jan. 2023, https://doi.org/10.48550/arXiv.2301.04246, accessed on 15.06.2024.

[162] ELEFTHERIA PAPAGEORGIOU, ET AL., *A Survey on the Use of Large Language Models (LLMs) in Fake News*, in «Future Internet», vol. 15, n. 8, 2023, https://doi.org/10.3390/fi16080298, accessed on 20.08.2024

In addition to the risk of encouraging disinformation campaigns, LLMs present a significant risk of misinformation, i.e. the unintentional dissemination of false and misleading information. As a matter of fact, given the architecture of these models and the phenomenon of mismatched generalisation,[163] they can generate inaccurate, false (hallucinations) or biased (gender, racial, etc.) outputs as discussed in section I.2.1. This leads to considerable implications in critical contexts such as medicine, law, or education. For instance, a model could combine accurate medical information with fictional elements, resulting in misdiagnosis suggestions. While, on one hand, «the critical challenge is that LLMs can be easily leveraged to generate deceptive misinformation at scale»,[164] on the other hand «LLMs bring promising opportunities for combating misinformation due to their profound world knowledge and strong reasoning abilities».[165] To increase LLMs' trustworthiness – i.e. the degree of reliability and security of these models in providing accurate, appropriate and useful answers – debiasing and dehallucinating techniques are being developed. Besides the debiasing approaches already tackled in section I.2.1, dehallucinating methods consist, amongst others, of integrating external knowledge bases, such as Wikipedia or other verified information databases, to improve the factuality of the answers.[166] Another interesting approach involves the utilisation of hybrid or neuro-symbolic models, which combine the power of machine learning with the rules of symbolic AI. This method consists in the construction of knowledge graphs for the model to verify the accuracy of generated outputs.[167]

[163] A. WEI, N. HAGHTALAB, J. STEINHARDT, *Jailbroken: How Does LLM Safety Training Fail?*, cit.

[164] CANYU CHEN, KAI SHU, *Combating Misinformation in the Age of LLMs: Opportunities and* Challenges, in *ArXiv*, 09 Nov. 2023, https://doi.org/10.48550/arXiv.2311.05656, accessed on 20.08.2024.

[165] *Ibidem*.

[166] ZICHAO LIN, ET AL., *Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models*, in «Artificial Intelligence Review», vol. 57, n. 243, 2024, https://doi.org/10.1007/s10462-024-10896-y, accessed on 20.08.2024.

[167] ALESSANDRO BRUNO, ET AL., *Insights into Classifying and Mitigating LLMs' Hallucinations*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.08117, accessed on 20.08.2024.

Legal responsibility, or liability, is a fundamental juridical concept referring to the obligation for an individual or an entity to be accountable for its actions and to compensate for any damage caused to third parties. Due to the autonomous and often unpredictable nature of AI systems, the establishment of liability for AI-generated content acquires greater complexity. If this content causes harm to other individuals, who should be held responsible?[168] «Many actors – designers, manufacturers, deployers, users – are involved in the chain of events and the training of an algorithm that can lead to a potential instance of harm; how to allocate fault among them is not very clear».[169] Not only can LLMs autonomously generate content potentially impacting both positively and negatively, but they can also make decisions based on complex algorithms and data that evolve over time. Their nature as 'black boxes' complicates the reconstruction of a clear and specific causal chain. In order to address these challenges, several approaches to liability have been proposed. One is strict liability, which suggests that agents and operators of AI models should be accountable for their actions and decisions, especially when their models have a significant impact on individuals or society,[170] regardless of fault. This approach, comparable to the one applied in many countries for defective products, would ensure that victims of harm caused by LLMs are able to obtain compensation without needing to prove negligence on the part of the provider or operator.[171] Another suggested approach is fault-based liability, wherein liability is only

[168] MATTHEW U. SCHERER, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, in «Harvard Journal of Law & Technology», vol. 29, n. 2, 2015, http://dx.doi.org/10.2139/ssrn.2609777, accessed on 20.08.2024.

[169] BEATRIZ BOTERO ARCILA, *Is it a platform? Is it a search engine? It's ChatGPT! The European liability regime for large language models*, in «Journal of Free Speech Law», vol. 3, 2023, https://ssrn.com/abstract=4539452, accessed on 20.08.2024, p. 473.

[170] NATALIA DÍAZ-RODRÍGUEZ, ET AL., *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*, in «Information Fusion», vol. 99, 2023, https://doi.org/10.1016/j.inffus.2023.101896, accessed on 20.08.2024.

[171] M. U. SCHERER, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, cit.

assigned when the provider, programmer or operator can be proven not to have exhibited an appropriate standard of care or to have acted negligently. This approach requires a detailed assessment of the conduct of the individuals involved in both the development and utilisation of LLM to determine whether they have met the required standard of care.[172] Lastly, a mixed approach has been suggested, according to which strict liability applies in some situations, such as in the case of highly risky products or in circumstances where human supervision is minimal, while fault-based liability is used in others.[173] «33% of firms view "liability for damage" as the top external obstacle to AI adoption, especially for LLMs, only rivalled by the "need for new laws", expressed by 29% of companies».[174] The need for a liability framework to protect the users of these systems without compromising innovation and development, as well as the numerous benefits demonstrated by the application of these models in different sectors, is compelling. The EU environment is developing a specific regulatory framework to address the challenges posed by generative AI and LLMs. Specifically, «the European Commission published a proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (the 'AI liability directive') in September 2022. The Commission proposes to complement and modernise the EU liability framework to introduce new rules specific to damages caused by AI systems».[175] This proposal introduces a fault-based liability procedure applicable to damage caused by AI generated content, directly complementing the regulatory framework of the AI Act. Furthermore, a proposal to update the Directive

[172] *Ibidem.*

[173] *Ibidem.*

[174] As reported in CLAUDIO NOVELLI, ET AL., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, in *ArXiv*, 15 Mar. 2024, https://doi.org/10.48550/arXiv.2401.07348, accessed on 20.08.2024.

[175] *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence* (COM/2022/496 final).

85/374/EEC on Product Liability[176] to specifically include AI systems is pending. It states that an AI system can be considered defective if it does not offer the level of safety a consumer might legitimately expect, considering all the circumstances, including statements by the provider and the knowledge gained after the product's release. However, there are several deficiencies in these directives when applied to generative AI, largely stemming from their reliance on the AI Act, which in turn appears to be inadequate to effectively regulate LLMs, because of the lack of certainty as to whether they fall under the high-risk categorised systems and which liability mechanisms are applicable. For instance, the classification based primarily on computational resources used for training (FLOPs) may not fully consider the complexity and impact of the models.[177] It is necessary to adopt a regulatory approach which precisely recognises LLMs' peculiarities, such as their ability to generate content autonomously and their potential impact on users and sectors.[178] Besides, the issue of accountability is closely linked to the concept of transparency and the urgent need to improve audit and monitoring mechanisms for AI systems. Without clear traceability and accountability of LLM decision-making, there is the risk of creating an 'accountability vacuum' where no one can be held responsible for the damage caused, undermining public trust in these technologies.

«What is interesting about LLMs is that they also raise concerns about 'positive' responsibility gaps: who, if anyone, can take credit for positive outputs?».[179] «Existing

[176] *Proposal for a Directive of the European Parliament and of the Council on liability for defective products* (COM/2022/495 final).

[177] C. NOVELLI, ET AL., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, cit.
Cfr. *AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice*, cit.

[178] N. DÍAZ-RODRÍGUEZ, ET AL., *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*, cit.

[179] S. PORSDAM MANN, ET AL. *Generative AI entails a credit–blame asymmetry*, in «Nature Machine Intelligence», vol. 5, 2023, pp. 472-475, https://doi.org/10.1038/s42256-023-00653-1, accessed on 20.08.2024.

copyright law does not recognize computer programs as authors, and hence, does not afford copyright protection to "work" created by computer programs».[180] Examples of AI-generated works include artworks such as *Portrait of Edmond Belamy*, originally created utilising a generative adversarial network (GAN)[181] and sold at Christie's auction for $ 432.500, far exceeding pre-sale estimates.[182] In Europe, intellectual property law is conventionally related to a human author and his or her own unique creativity. This anthropocentric perspective, although not explicitly stated in international treaties and EU law but only inferable from several regulations, establishes that in order to be entitled to copyright protection, a work must reflect the author's personality and intellectual capacity (something LLMs do not possess to date, despite being a complex philosophical debate).[183] Therefore, according to this perspective, only the person utilising the generative AI model (or the developers and provider of the model itself) could be the IP rights holder. For instance, according to the UK *Copyright, Designs and Patents Act* (CDPA) of 1988 «for "computer-generated" work, the "author" [...] is deemed to be the person who undertook "the arrangements necessary for the creation of the work." "Computer-generated" is defined as meaning that the work was "generated by computer in circumstances such that there is no human author of the work"».[184] However, the

[180] R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit., p. 148.

[181] It is a type of artificial neural network comprising two principal neural networks, 'competing' against each other (the generator and the discriminator). While the generator keeps improving to deceive the discriminator, the latter advances in order to detect falsehoods in the generator. The process proceeds until the generator succeeds in producing such realistic outputs the discriminator is unable to distinguish them from real data. The GANs are used in various fields, including image, music and video generation, music generation.

[182] CHRISTIE'S, *Obvious and the interface between art and artificial intelligence*, 12 Dec. 2018, https://www.christies.com/en/stories/a-collaboration-between-two-artists-one-human-one-a-machine-0cd01f4e232f4279a525a446d60d4cd1, accessed on 20.08.2024.

[183] C. NOVELLI, ET AL., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, cit.

[184] As cited in SIMON CHESTERMAN, *Good models borrow, great models steal: intellectual property rights and generative AI*, in «Policy and Society», 2024, https://doi.org/10.1093/polsoc/puae006, accessed on 20.08.2024.

question is still up for debate. In scenarios where a generative AI system it is but a mere instrument in the hands of a human inventor – and thus the creative, intellectual activity is only supported by artificial intelligence, yet remains a prerogative of the user who interacts with it – the current framework remains applicable.[185] For instance, an individual employing an LLM for stylistic suggestions, but actively writing the final text, could be granted copyright on the work created as it is a result of his or her intellectual creation. Specifically, the concept of 'intellectual creation' can be found in the European Court of Justice (CJEU) judgment in *Case C-5/08 Infopaq International A/S v Danske Dagblades Forening*. The dispute concerned the copyright and utilisation of newspaper article extracts by Infopaq, a Danish media monitoring company. The matter was brought before the CJEU to determine whether Infopaq's actions constituted an infringement of IP rights under Directive 2001/29/EC of the European Parliament and the Council, regarding the harmonisation of certain copyright and related rights within the information society (also known as the InfoSoc Directive). «It is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation. Words as such do not, therefore, constitute elements covered by the protection».[186] Nevertheless, when LLMs operate in a fundamentally autonomous manner, as the mere formulation of a prompt by a human creator is not sufficient to acknowledge a substantial contribution to the AI-generated output, or when the output is only subtly edited and not part of an actual intellectual creation by the user, copyright protection becomes more problematic.[187]

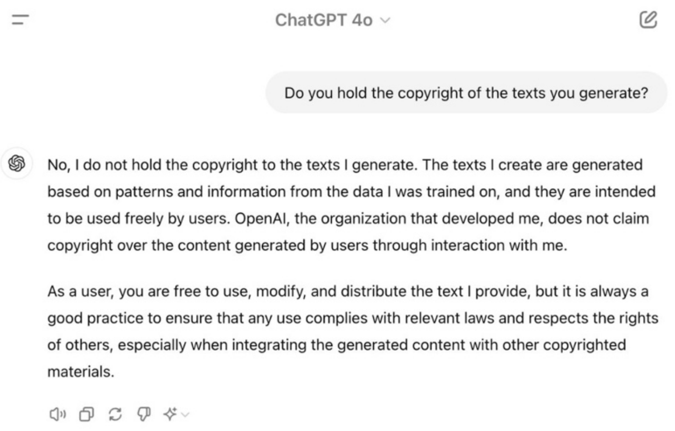
[185] C. NOVELLI, ET AL., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, cit.

[186] COURT OF JUSTICE OF THE EUROPEAN UNION, Judgment of 16 July 2009, *Infopaq International A/S v Danske Dagblades Forening, Case C-5/08*, paragraphs 46-47.

[187] *Ibidem.*

«Where human input does not reach a threshold of significance, the work remains authorless. To resolve such delineation issues, new models such as 'contributorship' for generated works and labelling duties should be explored».[188] This would embody the augmented intelligence and co-creation paradigm, providing an advantageous model for steering people in using LLMs to generate and share new ideas or contents, while remaining responsible in crucial aspects such as «meaning-making, imbuing of intention, showing creativity in prompt design and elaboration of initial outputs, careful editing, vetting, fact-checking, and other necessary contributions».[189] The AI Act does not contain any specific provisions regarding IP rights, and an ongoing legal debate is taking place on how intellectual property and patent laws could evolve to better deal with these challenges. While some argue that the existing legal framework can be applied without substantial changes, others, for instance, are suggesting a broader interpretation of the non-obviousness requirement in the case of patenting: it should incorporate the assessment of non-obviousness for an AI-assisted practitioner, thus taking into account the role of generative AI in the innovation process.[190] Worldwide, the position is heterogeneous: in 2021, the South African Patent Office granted a patent with an AI as inventor, setting a worldwide precedent.[191] As for OpenAI, the company specifically expressed it would not claim copyright of the content generated by ChatGPT (Figure 10).[192]

[188] S. P. MANN, ET AL. *Generative AI entails a credit–blame asymmetry*, cit.
[189] ID., ET AL., *AUTOGEN and the Ethics of Co-Creation with Personalized LLMs – Reply to the Commentaries*, in «The American Journal of Bioethics», vol. 24, n. 3, 2024, https://doi.org/10.1080/15265161.2024.2308175, accessed on 20.08.2024.
[190] ID., ET AL. *Generative AI entails a credit–blame asymmetry*, cit.
[191] ED CONLON, *DABUS: South Africa issues first-ever patent with AI inventor*, in «Managing IP», 29 Jul. 2021, https://www.managingip.com/article/2a5czh91g6c8zwxjcpla8/dabus-south-africa-issues-first-ever-patent-with-ai-inventor, accessed on 20.08.2024.
[192] OPENAI, *Terms of use*, https://openai.com/policies/row-terms-of-use/, accessed on 20.08.2024.



Do you hold the copyright of the texts you generate?

No, I do not hold the copyright to the texts I generate. The texts I create are generated based on patterns and information from the data I was trained on, and they are intended to be used freely by users. OpenAI, the organization that developed me, does not claim copyright over the content generated by users through interaction with me.

As a user, you are free to use, modify, and distribute the text I provide, but it is always a good practice to ensure that any use complies with relevant laws and respects the rights of others, especially when integrating the generated content with other copyrighted materials.

FIGURE 10*, Interaction with ChatGPT regarding the copyright of generated content.*

The intellectual property issue regarding the AI-generated outputs is not only a legal matter, but also an ethical one, as creativity is a widely debated idea without an unequivocal definition. According to the definition proposed by Margaret Boden, creativity implies the presence of three fundamental criteria with respect to a work: novelty, surprise and value, where «[n]ovelty refers to the dissimilarity between the produced artifact and other examples in its class»,[193] «[s]urprise instead refers to how much a stimulus disagrees with expectation», [194] and «[v]alue refers to utility, performance, and attractiveness. It is also related to both the quality of the output, and its acceptance by the society».[195] Notably, whilst LLMs are capable of producing novel content absent from the training data, and of a high quality, consequently satisfying the first and last criteria (novelty and value); it is the surprise criterion which most challenges the attribution of creativity to these models. In fact, being trained to follow pre-existing data patterns, they cannot be said to create surprising products. They possess combinatorial creativity, but hardly achieve the transformational creativity implying a

[193] GIORGIO FRANCESCHELLI, MIRCO MUSOLESI, *On the Creativity of Large Language Model*s, in *ArXiv*, 09 Jul. 2023, https://doi.org/10.48550/arXiv.2304.00008, accessed on 20.08.2024.
[194] *Ibidem.*
[195] *Ibidem*.

transformational change of thinking. Therefore, with the capabilities and architectures we know at present, LLMs can not be considered to be creative in the comprehensive sense of the term, and as defined by human creativity theories.[196]

The unresolved allocation of intellectual property and the integration of LLMs in various domains, particularly in scientific research and academic contexts, also presents significant ethical challenges for authorial attribution. Specifically, the fading of boundaries between human and machine authorship poses concerns about plagiarism, academic integrity, and authorship.[197] Authorship refers to the credit given for individuals who have made significant intellectual contributions to a piece of work, such as an academic paper, implying also accountability and intellectual property rights. As widely argued, the potential of LLMs in the process of co-creation of textual content is many: from assistance in drafting, to linguistic and stylistic revision, to proofreading, leading to their daily adoption also in academic practices, not without fears.[198] Nevertheless, should this contribution be made explicit, and if so how? For instance, after the publication of the paper *Can artificial intelligence help for scientific writing?*,[199] in which ChatGPT was cited as one of the authors, the publisher, Springer Nature, issued a correction by removing it, and justifying the action stating that LLMs do not fulfil their authorship criteria, since they cannot be effectively held accountable for the work produced.[200] Still, exploiting AI-generated content without proper citation can lead to serious ethical and

[196] *Ibidem.*

[197] JÜRGEN RUDOLPH, ET AL., *ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?*, in «Journal of Applied Learning & Teaching», vol. 6, no. 1, 2023, https://doi.org/10.37074/jalt.2023.6.1.9, accessed on 14.02.2024.

[198] *Ibidem.*

[199] MICHELE SALVAGNO, FABIO SILVIO TACCONE, ALBERTO GIOVANNI GERLI, *Correction to: Can artificial intelligence help for scientific writing?*, in «Critical Care», vol. 27, n. 99, 2023, h https://doi.org/10.1186/s13054-023-04390-0, accessed on 20.08.2024.

[200] *Submission Guidelines*, in *Springer Nature*, https://www.nature.com/commsbio/submit/submission-guidelines, accessed on 20.08.2024.

scholarly implications. It may lead readers to erroneously believe the author has actually written a work, thus not only compromising the author's honesty, but also being deceived about the real content source.[201] In an academic context, the outcomes of LLMs also undermine the traditional essay as assessment method. Students might employ them to produce an entire paper without conducting any real research, analysis, or critical reflection. A total prohibition of the usage of generative AI might not be the ideal solution, not only because the restriction of these tools would limit access to potentially enriching resources for the learning process (helping students to unravel complex concepts, develop innovative ideas, and assist them in the writing process), but also because in the contemporary workplace context, as noted in section I.2, the use of LLMs is increasingly widespread and in demand. Furthermore, AI-generated content may comprise biases and hallucinations of different kinds, and without human validation these are likely to foster misinformation and misinformation, as mentioned above. Rather, a proactive approach to address these issues entails, for instance, informing and educating students in academic contexts. First, there needs to be clarity and consistency on the part of institutions and educators regarding guidelines on the use of these technologies.[202] Secondly, educating individuals on how to ethically and responsibly employ these massive generative AI tools for research and writing, including understanding when to cite sources and how to avoid reliance on AI tools for critical thinking and analysis, can foster a safe and advantageous augmented intelligence paradigm.[203]

[201] BAIXIANG HUANG, CANYU CHEN, KAI SHU, *Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges*, in *ArXiv*, 16 Aug. 2024, https://doi.org/10.48550/arXiv.2408.08946, accessed on 20.08.2024.

[202] ALEXANDER J. CARROLL, JOSHUA BORYCZ, *Integrating large language models and generative artificial intelligence tools into information literacy instruction*, in «The Journal of Academic Librarianship», vol. 50, 2024, https://doi.org/10.1016/j.acalib.2024.102899, accessed on 20.08.2024.

[203] *Ibidem.*

I.4.4. Utilisation-related implications

The real-world contexts exploitation of LLMs, and more generally of generative AI, raises numerous ethical and legal implications: biases and fairness, privacy and data protection, misinformation and manipulation, accountability and transparency, intellectual property and copyright, economic and cultural impact. Without reiterating the implications already mentioned in the previous subsections, and which clearly have direct consequences in the practical utilisation of these models in people's everyday lives, this subsection will focus on the economic and social implications of the use of these technologies, not previously addressed.

Firstly, the deployment of LLMs is bringing about meaningful changes in the labour market by automating complex tasks which previously needed human intervention, such as customer service, content creation and data analysis, among others. In finance, for instance, LLMs can perform complex tasks such as analysing financial sentiments and forecasting stock movements.[204]

> While the industrial revolution mainly transformed physical work, foundation models are likely to transform tasks involving cognitive work, like content creation and communication. In general, since foundation models are intermediary assets that often possess strong generative capabilities, we envision that they will be able to augment humans in many creative settings, rather than replace humans as there are still significant limitations in using these models stand-alone for open-ended generative tasks.[205]

[204] ZHIYU ZOEY CHEN, ET AL., *A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law*, in *ArXiv*, 02 May 2024, accessed on 21.08.2024.

[205] R. BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit., p. 150.

Despite the processes automation leading to several productivity and efficiency advantages, it could also result in job displacement in many sectors,[206] manifested not only in a loss of employment, but also in a worsening wage disparity between highly qualified and less qualified workers.[207] For this reason, it is essential to foster in-depth literacy programmes at all levels, both academic and post-academic, such as the ones proposed by the Italian government.[208] Nevertheless, while job displacement is a major concern, in parallel new opportunities are arising in emergent fields, for instance the management, supervision, regulation and development of LLMs themselves.[209] A further economic and social implication of adopting LLMs is the huge demand for computing resources and specialised skills, rendering their access expensive and often limited to large technology companies, research institutions or well-funded governments.[210] The centralisation of decision-making rights and power may ultimately lead to an imbalance between the economically and socially more fragile: 'The Turing Trap'.[211]

Inequality of access is not only economic, but also geographical and linguistic: as addressed in Section I.3.1, most LLMs are trained on dominant language and cultural data (predominantly English), reducing their effectiveness for minority languages or less represented cultural contexts, further contributing to the marginalisation of non-dominant languages and cultures.[212] For instance, models such as GPT-3 and GPT-4 tend to

[206] KASSYM-JOMART TOKAYEV, *Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns*, in «International Journal of Social Analytics», vol. 8, n. 3, 2023, https://norislab.com/index.php/ijsa/article/view/42, accessed on 20.08.2024.

[207] As cited in TYNA ELOUNDOU, ET AL., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, in *ArXiv*, 21 Aug. 2023, https://doi.org/10.48550/arXiv.2303.10130, accessed on 20.08.2024.

[208] *Strategia italiana per l'intelligenza artificiale 2024-2026*, cit.

[209] T. ELOUNDOU, ET AL., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, cit.

[210] *Ibidem.*

[211] As cited in R, BOMMASANI, ET AL., *On the Opportunities and Risks of Foundation Models*, cit., p. 151

[212] SHAILY BHATT, *Extrinsic Evaluation of Cultural Competence in Large Language Models*, in *ArXiv*, 19 Jun. 2024, https://doi.org/10.48550/arXiv.2406.11565, accessed on 21.08.2024.

perform worse in languages with non-Latin scripts or resource-limited languages such as Arabic, Greek, Hindi, Turkish, Vietnamese and Chinese.[213] This under-representation has wide-ranging implications which can affect several aspects of society. On an individual level, minority culture users are likely not to see their values and experiences reflected, which could result in a decrease in the inclusivity and accessibility of LLMs-based technologies. Moreover, on a broader level, the widespread use of these models could reinforce cultural stereotypes and perpetuate cultural homogenisation, contributing to the marginalisation of already under-represented languages and cultures.[214] To address these problems, some researchers propose improving the datasets used to train LLMs by integrating more diverse cultural data, such as those collected by international surveys like the World Values Survey, to ensure that models have the possibility to learn a greater variety of cultural and linguistic values.[215] Nevertheless, it is essential that all stakeholders concerned recognise the importance of fairly representing all cultures and languages to ensure a more inclusive and representative future in generative AI.

Lastly, another significant social and psychological concern is the over-reliance on LLMs' technologies for tasks such as writing, problem-solving and decision-making. «The consequences of over-reliance on LLMs can be multifaceted. In the context of decision-making, whether in corporate settings or public policy, over-dependency on automated recommendations can potentially stifle human creativity, intuition, and ethical considerations that a machine model cannot encapsulate».[216] For instance, people may

[213] WENHAO ZHU, ET AL., *Extrapolating Large Language Models to Non-English by Aligning Languages*, in *ArXiv*, 09 Oct. 2023, https://doi.org/10.48550/arXiv.2308.04948, accessed on 21.08.2024.
[214] CHENG LI, ET AL., *CultureLLM: Incorporating Cultural Differences into Large Language Models*, in *ArXiv*, 09 Feb. 2024, https://doi.org/10.48550/arXiv.2402.10946, accessed on 21.08.2024.
[215] *Ibidem*.
[216] K.-J. TOKAYEV, *Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns*, cit., p. 25.

over-rely on the research and writing skills of these models, compromising the development of skills such as critical thinking and effective research. Or, they could fail to validate the generated information, by accepting the output as definitive without further human verification (automation bias),[217] a circumstance which appears highly concerning as it would lead to the spread of misinformation, disinformation or unacknowledged bias.[218]

It is only human oversight that can ensure that LLMs are used as tools assisting decision-making and creative activities, rather than replacing human judgement. «Tackling the issue of over-reliance involves a balanced approach that integrates LLMs into existing systems and processes while maintaining human oversight. Training programs can educate users about the limitations and best practices of using LLMs, encouraging a more informed and critical approach».[219]

[217] S. E. HUBER, ET AL., *Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning*, cit.

[218] UDARA PIYASENA LIYANAGE, NIMNAKA DILSHAN RANAWEERA, *Ethical Considerations and Potential Risks in the Deployment of Large Language Models in Diverse Societal Contexts*, in «Journal of Computational Social Dynamics», vol. 8, n. 11, 2023, https://vectoral.org/index.php/JCSD/article/view/49, accessed on 21.08.2024.

[219] K.-J. TOKAYEV, *Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns*, cit., p. 26.

SECOND CHAPTER

INTERACTING WITH MACHINES: HUMAN-AI INTERACTION

"We shape our tools and, thereafter, our tools shape us."
FATHER JOHN CULKIN

"Our intelligence is what makes us human, and AI is an extension of that quality."
YANN LECUN

II.1. Definition and overview

Historically, machines and computers have not been developed primarily for their intrinsic value or as ends in themselves, but rather as instruments to support human endeavours, aiding in solving problems and carrying out heterogeneous tasks, and fostering scientific, cultural, and industrial progress. As noted, «[t]he term "intelligence amplification" seems applicable to our goal of augmenting the human intellect in that the entity to be produced will exhibit more of what can be called intelligence than an unaided human could».[220] For as long as there have been computers, therefore, people have

[220] DOUGLAS CARL ENGELBART, *Augmenting Human Intellect: a Conceptual Framework*, Menlo Park, Stanford Research Institute, 1962.

interacted with them in different modalities, with varying intentions and degrees of knowledge.

Human-Computer Interaction (HCI) is an interdisciplinary field specifically concerned with investigating the interactions between humans and technology, by integrating perspectives from computer science, cognitive psychology, design and visual arts, ergonomics, linguistics, sociology, ethics, and more.[221] The purpose of research is not purely theoretical, but rather pragmatic. The intention is to investigate «to what extent computers are or are not developed for successful interaction with human beings»[222] so as to consistently improve them, enhancing their sophistication and suitability for human needs.

The origins of HCI can be traced back to the 1960s, when initial studies endeavoured to facilitate the employment of computers by non-technical users through more versatile and user-friendly interfaces and hardware. This necessity gave rise to the establishment of design paradigms and theories continuing to shape the field today,[223] most notably the concept of Human-Centred Design (HCD). HCD is a methodological approach prioritising the needs, capabilities and values of users in relation to a system, whether hardware or software, throughout the entire technological development cycle. Differently from traditional design frameworks, which often focus predominantly on technical or economic efficiency, HCD strives to align product development with key

[221] *The Evolution of Human-Computer Interaction: A Review of the Past and Future Directions*, in *Association of Human-Computer Interaction*, https://www.hci.org.uk/article/the-evolution-of-human-computer-interaction-a-review-of-the-past-and-future-directions/, accessed on 20.08.2024.

[222] AHMED AWAD E. AHMED, Employee Surveillance Based on Free Text Detection of Keystroke Dynamics, in MANISH GUPTA, RAJ SHARMAN, *Handbook of Research on Social and Organizational Liabilities in Information Security*, Hershey, IGI Global, 2009.

[223] ALAN DIX, *Human-Computer Interaction*, in LING LIU, M. TAMER ÖZSU, *Encyclopedia of Database Systems*, Boston, Springer, 2009, https://doi.org/10.1007/978-0-387-39940-9_192, accessed on 20.08.2024.

principles of usability, accessibility, engagement, and intuitiveness.[224] For instance, a breakthrough in personal computing has been the introduction of the desktop metaphor (imitating a physical office desk, with files, folders and documents displayed as icons, as well as windows and menus) within graphical user interfaces (GUIs), as it provided a familiar and intuitive framework for users to interact with digital environments. This change, made popular by systems such as the Xerox Star by Xerox PARC and Macintosh by Apple Inc. in the 1980s, enabled people to interact with computers without having to understand complex command-line interfaces (CLIs).[225] HCD is underpinned by fundamental principles guiding the design and deployment of technology, such as continuous user engagement through participatory design, an iterative process involving the refinement of solutions based on cyclical user feedback, and a deep recognition of the significance of empathy and human diversity for technology to be widely accessible, inclusive, and non-discriminatory. Achieving these aims necessitates interdisciplinary cooperation between engineers, designers, sociologists, ethicists, and end-users alike.[226] One of the most significant contributions to the field of HCD was the one of Donald Norman, author of the book *The Design of Everyday Things* in 1988, which still today serves as a fundamental reference for practitioners in the field, as it contains the theorisations of affordances, signifiers and feedback within design. Specifically, affordances refer to the properties of objects indicating their potential usages. They are not inherent properties, rather they represent the relationship between users and objects

[224] JAN AUERNHAMMER, *Human-centered AI: The role of Human-centered Design Research in the development of AI*, in STELLA BOESS, MING CHEUNG, REBECCA CAIN (edited by), «Proceedings of DRS2020 International Conference», 2020, https://doi.org/10.21606/drs.2020.282, accessed on 01.09.2024.

[225] VLADIMIR L. AVERBUKH, *Sources of Computer Metaphors for Visualization and Human-Computer Interaction*, in *IntechOpen*, 17 Jun. 2020, https://doi.org/10.5772/intechopen.89973, accessed on 01.09.2024.

[226] WEI XU, ZAIFENG GAO, *Enabling Human-Centered AI: A Methodological Perspective*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.06703, accessed on 01.09.2024.

themselves. On the other hand, signifiers are deliberate cues embedded in design for communicating how to effectively utilise the object, making the functionality explicit. Lastly, feedback ensures users are informed immediately of the consequences of their actions, a vital aspect of interaction design. These three elements form the backbones of interaction design, emphasising usability, accessibility, and user empowerment.[227]

As HCI progressed, in the 1960s the first automation systems emerged, known as 'expert systems', designed to replicate human decision-making through symbolic approaches and which – despite their limitations in terms of learning and generalisation capabilities – laid the foundation for subsequent evolutions in interactions with artificial intelligence.[228] Notably, the chatbot or conversational agent ELIZA (referred to in Chapter I.1) was conceived specifically as an experimental venture to investigate human-machine interactions. Weizenbaum, ELIZA's developer, was first and foremost surprised by how intensely users responded to the agent's outputs. Even those having technical expertise and a comprehension of its programming and internal functioning, found themselves interacting with the system as if it possessed actual knowledge and intentions, much like a human interlocutor. This behaviour was first identified as an example of the cognitive bias leading people to ascribe human-like qualities to a machine based solely on surface-level and seemingly intelligent interactions.[229] This psychological phenomenon came to be known as the ELIZA effect, and it holds significant ethical, social, and psychological implications. As a matter of fact, the unrealistic attribution of in-depth cognitive faculties to computers raises concerns about how technology can be

[227] DONALD NORMAN, *The Design of Everyday Things*, Cambridge, MA, The MIT Press, 2013.

[228] BABAK ABEDIN, ET AL., *Designing and Managing Human-AI Interaction*, in «Information System Frontiers», vol. 24, 2022, pp. 691-697, https://doi.org/10.1007/s10796-022-10313-1, accessed on 23.08.2024.

[229] J. WEIZENBAUM, *ELIZA - a computer program for the study of natural language communication between man and machine*, cit.

misperceived and uncritically relied upon, leading to over-reliance on computer-generated outputs regardless of their lack of depth and genuine understanding.[230] Specifically, Weizenbaum's critiques turn out to be of particular relevance in contemporary discussions regarding AI ethics, where the over-reliance in systems empowered by algorithms increasingly advanced and socially, economically, pragmatically pervasive could lead to problematic consequences in decision-making and human autonomy.

Since the 1990s, as neuroscience and cognitive sciences evolved, theories pertaining to these fields have also been applied to HCI, encouraging the development of more sophisticated and user-centred design methodologies,[231] which then became the prevailing paradigm, focusing on the efficiency, accessibility and intuitiveness of user interfaces. In recent years, Human-Computer Interaction has been heavily impacted by several shortcomings, including the growing use of mobile devices and touch screens, (aligned with Mark Weiser's[232] vision of ubiquitous computing in which information processing is entirely integrated within everyday objects), increasingly frequent online interactions, the popularity of virtual reality (VR) and augmented reality (AR), and the sophistication of artificial intelligence algorithms. The latter has been so pervasive as to have generated a neologism to be referred to, namely Human-AI Interaction (HAII, HAI or HAX).

The history of HAII is deeply and intrinsically intertwined with the development of intelligent systems, as previously explored in the first chapter. With the exponential

[230] DAVID M. BERRY, *The Limits of Computation*, in «Weizenbaum Journal of the Digital Society» vol. 3, n. 3, 2023, https://doi.org/10.34669/WI.WJDS/3.3.2, accessed on 23.08.2024.
[231] *Ibidem*.
[232] Mark D. Weiser (Harvey, 1952 – Palo Alto, 1999), was a computer scientist and CTO at Xerox PARC, known for being a pioneer of technological innovation.

advances in the efficiency of deep neural networks, algorithmic design, and the volume of data used for training, interactions with machines have undergone a dramatic transformation. Where once computers were designed merely to automate specific tasks, people now engage with highly sophisticated systems capable of being active and creative partners. At the heart of this paradigm shift lies the way in which intelligence itself is being redefined through this interaction, moving from human-focused to human-integrated processes. Essentially, HAII is characterised by the in-depth capabilities of AI systems, and specifically of GAI, to engage in interactions beyond predefined rules or straightforward process automation tasks. These technologies now participate in dialogic exchanges, decision-making and even creative processes, pushing the boundaries of human-machine collaboration. This evolution has been largely driven by the improvements in natural language processing (NLP) capabilities, which has played a crucial role in enabling more natural and meaningful interactions. The majority of human interactions are, as a matter of fact, mediated through the use of natural language. Specifically, not only is NLP a central catalyst in the evolution of user interfaces, but it also empowers machines to iteratively comprehend and refine the context, intentions and subtleties of user input. This marks a departure from the traditional input-action model (with linear and largely predictable interactions: a person provided a command, and the machine precisely executed it), ushering a new era of agent-based models, enabling continuous and natural dialogues where AI can learn from past interactions and adapt to user preferences, thus enhancing the fluidity and depth of human-AI interactions.[233]

[233] JIAYIANG LI, JIALE LI, YUNSHENG SU, *A Map of Exploring Human Interaction Patterns with LLM: Insights into Collaboration and Creativity*, in HELMUT DEGEN, STAVROULA NTOA (edited by), *Artificial Intelligence in HCI*, vol. 14735, Cham, Springer, 2024, https://doi.org/10.1007/978-3-031-60615-1_5, accessed on 02.09.2024.

> Recently, the development of Large Language Models (LLMs) has revolutionized the capabilities of AI systems. These models possess the ability to comprehend and generate human-like text, enabling them to engage in sophisticated conversations, generate content, and even perform tasks that once seemed beyond the reach of machines. As a result, the way we interact with technology and each other – an established field called "Human-AI Interaction" and have been studied for over a decade – is undergoing a profound transformation.[234]

There is one specific milestone date, the 30th of November 2022, on which LLMs interaction has undergone a major disruption. Namely, the worldwide release of ChatGPT, the chatbot developed by OpenAI. «ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public»,[235] «ChatGPT is more advanced than any other chatbot available for public interaction, and many observers say it represents a step change in the industry. "Talking" to it can feel bewitching»,[236] «[t]he tool quickly went viral. On Monday, Open AI's co-founder Sam Altman, a prominent Silicon Valley investor, said on Twitter [X] that ChatGPT crossed one million users»,[237] wrote major newspapers within days of the release. The popularity was primarily driven by its distinctively user-friendly interface. As opposed to early LLMs that were only accessible to computer programmers through complex command-line interfaces (CLIs) or application programming interfaces (APIs) – by employing programming libraries such as TensorFlow or PyTorch on Python – or that were integrated into specific

[234] Diyi Yang, Sherry Tongshuang Wu, Marti A. Hearst, *Human-AI Interaction in the Age of LLM*s, in «Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies», vol. 5, pp. 34-38.

[235] Kevin Roose, *The Brilliance and Weirdness of ChatGPT*, in «The New York Times», 1 Mar. 2023, https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html, accessed on 15.02.2024.

[236] Billy Perrigo, *AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits*, in «Time», 05 Dec. 2022, https://time.com/6238781/chatbot-chatgpt-ai-interview/, accessed on 20.03.2024.

[237] Samantha Murphy Kelly, *This AI chatbot is dominating social media with its frighteningly good essays*, in «CNN Business», https://edition.cnn.com/2022/12/05/tech/chatgpt-trnd/index.html, accessed on 04.07.2024.

applications, ChatGPT was engineered with both simplicity and broad approachability in mind. Anyone with an email address has the opportunity to generate text for a wide range of tasks, without needing prior knowledge of programming languages, machine learning techniques, or the complex inner workings of a transformer. This democratisation extended across different users: students, professionals from various fields, such as doctors, lawyers, and educators, as well as creative and marketing practitioners. By abstracting from technical complexity, current LLMs allow users to exploit the in-depth capabilities of GAI by means of 'simple' natural language commands.

In addition, whereas originally prompts were only textual instructions, multimodal large language (MMLLM) models have soon allowed for deeper engagement with different input and output modalities, such as text, images, videos, and audios. Underlying these models is the concept of multimodal feature fusion mechanism, which refers to the method by which inputs from different modalities are integrated within a coherent system to create a unified representation. The fundamental architecture often consists of a basic pre-trained language model combined with additional modality-specific encoders. They process inputs using mechanisms such as cross-attention (in which one modality, such as text, is treated as a query, while the other modality, such as image features, is treated as keys and values) or feature projections (in which features from different modalities are projected into a shared latent space through linear transformations).[238] For instance, in the health sector, models such as LLaVA-Med integrate visual and textual data to improve diagnostic capabilities by combining pictures (i.e. medical CT scans) with the corresponding diagnosis description.[239] Multimodal LLMs are not only capable of

[238] DUZHEN ZHANG, ET Al., *MM-LLMs: Recent Advances in MultiModal Large Language Models*, in *ArXiv*, 28 May 2024, https://doi.org/10.48550/arXiv.2401.13601, accessed on 02.09.2024.

[239] CHUNYUAN LI, ET AL., *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*, in *ArXiv*, 01 Jun. 2023, https://doi.org/10.48550/arXiv.2306.00890, accessed on 02.09.2024.

comprehending and generating sophisticated images, but have also experienced significant improvements in understanding and generating audio and video. In September 2023, OpenAI introduced the opportunity for users to interact verbally with the model GPT-4, enabling the LLM-based chatbot to respond with synthetic voice, leveraging both text-to-speech (TTS) and automatic speech recognition (ASR) technologies, ultimately increasing accessibility and usefulness in real-world applications. The system is not only designed to realistically mimic human dialogue in terms of fluency, tone, and cadence, but also to maintain contextual flow within the conversation.[240] As of September 2024, major tech companies are greatly investing in the production of multimodal LLMs, not only OpenAI's GPT-4 and GPT-4o and Anthropic PBC's Claude, but also Google LLC's Gemini and Meta Platforms Inc.'s LLaMA 2, hoping they will result in rapid advancements towards general artificial intelligence (AGI), i.e. a hypothetical chronological future point at which technological progress surpasses human cognition and the ability to forecast its consequences.

However, such developments blur the lines between traditional human-computer interfaces and more immersive human-AI environments, pushing the boundaries of what is considered 'interaction' in a technical and social sense, and leading to overly misaligned expectations or undue apprehensions regarding the dystopian impacts of these systems. Specifically, people are naturally inclined to anthropomorphise non-human entities because of the wider tendency to seek social cues in communication and reacting to interactional behaviour as a sign of personality. On the one hand, reliance on and trust in such systems facilitates interactions and engagement, encouraging the emergence of

[240] REEM TEMSAH, ET AL., *Healthcare's New Horizon With ChatGPT's Voice and Vision Capabilities: A Leap Beyond Text*, in «Cureus», vol. 15, 2023, https://doi.org/10.7759/Fcureus.47469, accessed on 02.09.2024.

positive emotions and satisfaction. On the other hand, not only the risk of anthropomorphising GAI bear with it over-reliance repercussions and unaware delegation of decision-making responsibility to these systems, but users may also develop emotional attachments similar to the ones developed during human-to-human interactions, especially when dealing with LLMs that consistently and credibly simulate emotional intelligence.[241] For instance, Replika, a GAI system allowing people to create virtual companions capable of writing, calling, and sending voice messages, and which currently holds twenty million users worldwide, is raising troubling issues. Its chatbots have been known for their ability to engage in emotionally charged conversations, sometimes acting as a source of comfort, and ultimately resulting in situations of emotional dependence on them. «They found that some users were forming maladaptive bonds with their virtual companions, centering the needs of the AI system above their own and wanting to become the center of attention of that system».[242] Ethical and legal constraints regarding the long-term effects and damages of this possible emotional dependency, such as a decrease in human social connections or the emergence of unrealistic expectations of AI's capabilities, are growing and still the subject of debate, with researchers urging caution when using these systems in sensitive contexts such as psychological therapy. While, through affective computing, integrating emotional intelligence within computational systems is essential for more nuanced and natural interactions, at the same time, the risk of over-anthropomorphising by design these models is likely to further undermine the

[241] LUOMA KE, ET AL., *Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review*, in *ArXiv*, 16 Mar. 2024, https://doi.org/10.48550/arXiv.2401.01519, accessed on 02.09.2024.

[242] CLAIRE BOINE, *Emotional Attachment to AI Companions and European Law*, in *MIT Case Studies in Social and Ethical Responsibilities of Computing*, 27 Feb. 2023, https://doi.org/10.21428/2c646de5.db67ec7f, accessed on 02.09.2024.

already established user predisposition towards perceiving AI as a sentient and conscious entity.

Furthermore, the complexity of the neural networks powering these AI systems, their inherent limitations, as previously discussed (i.e. transparency, explainability, lack of semantics, privacy and intellectual property constraints, and so on) and their deployment in critical areas raise several dangerous issues for society and individuals:

> While AI technology has brought in many benefits to humans, it is having a profound impact on people's work and lives. [...] Many AI professionals are primarily dedicated to studying algorithms, rather than providing useful AI systems to meet user needs, resulting in the failure of many AI systems. Specifically, the AI Incident Database has collected more than 1000 AI related accidents, such as an autonomous car killing a pedestrian, a trading algorithm causing a market "flash crash" where billions of dollars transfer between parties, and a facial recognition system causing an innocent person to be arrested.[243]

These ethical and legal implications highlight the importance of adopting a human-centred approach in the design, development, use, and implementation of artificial intelligence. This paradigm, i.e. human-centred AI (HCAI), concentrates on maintaining the user at the focus of design, development, deployment, and implementation processes. It can be regarded «as a Second Copernican Revolution, promoting the idea of putting humans at the center with algorithms with AI orbiting nearby, instead of putting algorithms and AI at the center».[244] HCAI aims towards systems that strengthen human cognitive and creative capacities, leading to fruitful cooperation, and not to a substitution or marginalisation of human competences, originality and decision-making. Specifically,

[243] D. YANG, S. TONGSHUANG WU, M. A. HEARST, *Human-AI Interaction in the Age of LLM*s, cit.
[244] As cited in W. XU, Z. GAO, *Enabling Human-Centered AI: A Methodological Perspective*, cit.

«HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy. These systems also support human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation».[245] One major challenge concerning the development and deployment of human-centric AI is striking a balance between the necessity to scale the inherent technical capabilities of the models with human ethical and legal concerns. In this respect, the European Union not only established the AI Act regulatory framework, whose main concern is the safeguarding of democratic and humanistic values through the control of transparency, explicability and human ability to comprehend and control AI systems; but on 8 April 2019, it also presented the *Ethics Guidelines for Trustworthy Artificial Intelligence*.[246] Notably, these guidelines list a set of 7 essential requirements for the development of trustworthy AI systems:

1. Human agency and oversight
   Including fundamental rights, human agency and human oversight
2. Technical robustness and safety
   Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. Privacy and data governance
   Including respect for privacy, quality and integrity of data, and access to data
4. Transparency
   Including traceability, explainability and communication
5. Diversity, non-discrimination and fairness
   Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

[245] BEN SHNEIDERMAN, *Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy*, in «IUI '21 Companion: Companion Proceedings of the 26th International Conference on Intelligent User Interfaces», 2021, https://doi.org/10.1145/3397482.3453994, accessed on 30.08.2024.

[246] EUROPEAN COMMISSION, *Ethics Guidelines for Trustworthy AI*, 8 Apr. 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, accessed on 20.05.2024.

6. Societal and environmental wellbeing
   Including sustainability and environmental friendliness, social impact, society and democracy
7. Accountability
   Including auditability, minimisation and reporting of negative impact, trade-offs and redress.[247]

While the majority of these aspects have been addressed in Chapter I, the importance of maintaining human beings as the focus of attention within every phase of the design, development and utilisation of AI so as to foster meaningful interactions is the ultimate goal of HAII. Notably, this is reachable also through the integration of human oversight, a supervision acting as a governance mechanism to manage the risks associated with AI, within interactions with these technologies. Several studies have demonstrated how users interacting with advanced models of GAI are likely to over-trust, especially when they display both fluency and confidence in the outputs generated. A recent experiment on answering medical questions via LLM assistance revealed when people were presented with answers exhibiting high confidence and no uncertainty, they were far more likely to accept incorrect answers compared to when such uncertainty was clearly stated through natural language.[248] Human oversight is not only crucial for monitoring and eventually overriding or reversing AI-generated decisions and outputs, but also for critically analysing the outputs in terms of hallucinations, biases and common-sense nuances AI struggles to deal with. This also applies to interactions with LLMs, such as those underlying ChatGPT. The flowchart designed by Aleksandr

[247] *Ibidem.*

[248] SUNNIE S. Y. KIM, ET AL., *"I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust*, in *ArXiv*, 15 May 2024, https://doi.org/10.48550/arXiv.2405.00623, accessed on 29.08.2024.

Tiulkanov concisely exemplifies the concept whilst helping users to make informed and safe decisions:

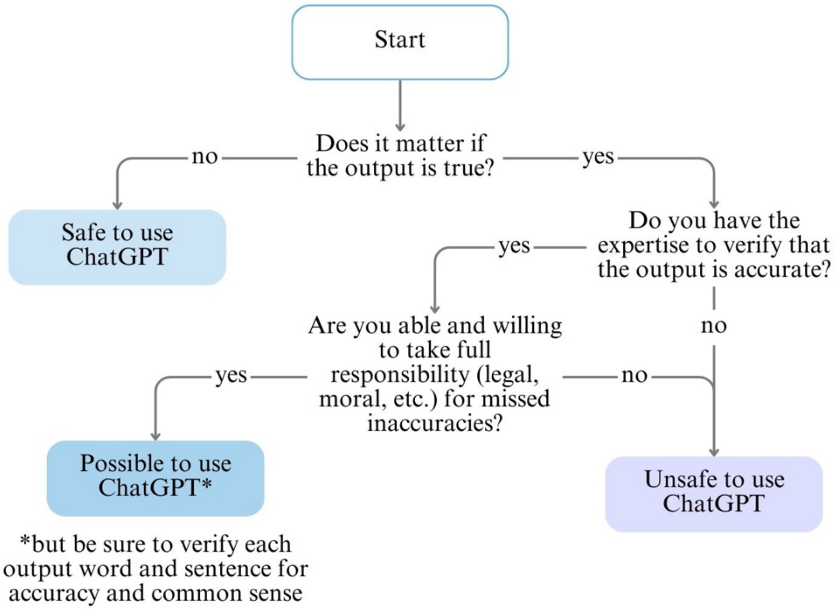

Start
Does it matter if the output is true?
no
yes
Safe to use ChatGPT
Do you have the expertise to verify that the output is accurate?
yes
no
Are you able and willing to take full responsibility (legal, moral, etc.) for missed inaccuracies?
yes
no
Possible to use ChatGPT*
Unsafe to use ChatGPT
*but be sure to verify each output word and sentence for accuracy and common sense

FIGURE 11*, When is it safe to use ChatGPT? Flowchart designed by Aleksandr Tiulkanov.*[249]

However, concerns about the effectiveness of this oversight remain, particularly the risk of people merely passively validating results without real substantive scrutiny: «[f]or meaningful human control, the decision-making system must be able to both track relevant moral reasons and trace back to an individual along the chain who is aware and accepting of the responsibility».[250] One operational model allowing implementation of human oversight is the human-in-the-loop approach (HITL), which – within the

[249] ALEKSANDR TIULKANOV, «A simple algorithm to decide whether to use ChatGPT, based on my recent article», in *X*, https://x.com/shadbush/status/1616007675145240576, accessed on 02.04.2024.
[250] LEILA METHNANI, ET AL., *Let Me Take Over: Variable Autonomy for Meaningful Human Control*, in «Frontiers Artificial Intelligence», vol. 4, 2021, https://doi.org/10.3389/frai.2021.737072, accessed on 24.07.2024.

discussion on human-centred artificial intelligence – relates to the development of systems in which human input is integral to the various stages of the process. This is crucial in applications where human judgement is paramount and where the complexity or uncertainty of the system requires critical thinking, in consideration of the intrinsic limitations the architecture of deep learning models entails. «Introducing human intelligence to the loop of intelligence systems can realize a close coupling between the analysis-response advanced cognitive mechanisms in fuzzy and uncertain problems and the intelligent systems of a machine».[251] This seamless integration, which therefore displaces the chance of complete replacement and deprivation of human decision-making faculties, is embodied in a paradigm of augmented intelligence. It stands as a synergetic, respectful and effective collaboration between people and AI, where the goal is to leverage mutual assets to create coherent and relevant outputs, as opposed to concentrating on technological development of systems detached from their relationship with people. This is beyond mere automation. It is about creating systems that can interpret, predict and reason in tandem with humans, enriching decision-making.[252]

> Human-in-the-loop (HITL) hybrid-augmented intelligence is defined as an intelligent model that requires human interaction. In this type of intelligent system, human is always part of the system and consequently influences the outcome in such a way that human gives further judgment if a low confident result is given by a computer. HITL hybrid-augmented intelligence also readily allows for addressing problems and requirements that may not be easily trained or classified by machine learning.[253]

[251] NAN-NING ZHENG, ET AL., *Hybrid-augmented intelligence: collaboration and cognition*, in «Frontiers of Information Technology & Electronic Engineering», vol. 18, 2017, pp..153-179, https://doi.org/10.1631/FITEE.1700053, accessed on 10.04.2024.

[252] *Ibidem.*

[253] *Ibidem.*

For instance, in the corporate sector, AI has been successfully implemented to provide analyses supporting complex decisional processes within management boards. For strategic planning, such systems can analyse market trends and consumer behaviour to inform product development and marketing strategy choices.[254] Human-AI teaming is a fast-advancing field in terms of both academic research and real-world implementations. The AI-enabled agents are in this case autonomous teammates capable of cooperating with human partners to achieve certain objectives.[255]

The future of human-AI interaction is complicated and so rapidly moving forward as to present several opportunities and challenges, particularly with the integration of these systems into everyday life. As the relationship between these new, powerful technologies deepens, interdisciplinary research requires focusing on a number of dimensions, such as ethical concerns, transparency of algorithms, and collaborative efficiency in order to promote a technological innovation beneficial to individuals and society as a whole. As AI increasingly gains the capacity to act independently, the need for structures and frameworks to ensure ethical development and behaviour of the models and their alignment with human values increases. HCAI is establishing itself as a solution, fostering approaches and methodologies prioritising human welfare and ethical standards.

[254] MANAL AHDADOU, ABDELLAH AAJLY, MOHAMED TAHROUCH, *Unlocking the potential of augmented intelligence: a discussion on its role in boardroom decision-making*, in «International Journal of Disclosure and Governance», vol. 21, 2024, pp. 433-446, https://doi.org/10.1057/s41310-023-00207-2, accessed on 03.09..2024.

[255] W. XU, Z. GAO,

THIRD CHAPTER

COMMUNICATING WITH MACHINES: PROMPT ENGINEERING

"Speak clearly, if you speak at all; carve every word before you let it fall."
OLIVER WENDELL HOLMES SR.

"The key to artificial intelligence has always been the representation."
JEFF HAWKINS

III.1. Definition and principal techniques

Among the emerging abilities that LLMs are shown to exhibit is few-shot learning[256] – also known as in-context learning. It allows these models to generalise and perform complex tasks based on a limited number of examples provided within the context of a single query.[257] As already covered, while these models are not inherently designed to perform specific jobs, they can be 'programmed' in a more targeted manner through fine-tuning, adjusting the model architecture, or through prompt optimisation,

[256] T. B. BROWN, ET AL., *Language Models are Few-Shot Learner*s, cit.

[257] PENGFEI LIU, ET AL., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, in *ArXiv*, 28 Jul. 2021, https://doi.org/10.48550/arXiv.2107.13586, accessed on 15.07.2024.

improving the results achieved at the level of an individual interaction.[258] Prompts are natural language instructions provided as of textual inputs to generative AI systems to elicit a specific action, such as text generation in the case of LLMs, or image creation in the case of text-to-image generative models.[259] Prompt engineering, on the other hand, concerns the design and specialisation of these prompts in order to enhance the quality of the output generated by the model. This expertise is becoming increasingly important for performing effective and targeted interactions with LLMs, especially in scenarios where new data retraining is either limited or impractical. In fact, it does not require altering the model's architecture or the employment of extensive datasets, thereby significantly reducing the computational costs and effort. [260]

> While prompting can appear as easy as instructing a human, crafting effective and generalizable prompt strategies is a challenging task. How a prompt or a prompt strategy directly impacts model outputs, and how prompts modify LLMs' billions of parameters during re-training, are both active areas of NLP research. […] Even for NLP experts, prompt engineering requires extensive trial and error, iteratively experimenting and assessing the effects of various prompt strategies on concrete input-output pairs, before assessing them more systematically on large datasets.[261]

Several studies have demonstrated that the effectiveness of AI-human interactions, and, consequently, the quality and originality of the outputs generated, is intrinsically linked to the quality of the prompts provided by users. As noted, «[o]ne vital skill of the 21st century could be effectively talking to machines. And for now, that process involves

[258] JULES WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, in *ArXiv*, 21 Feb. 2023, https://doi.org/10.48550/arXiv.2302.11382, accessed on 13.04.2024.

[259] For instance, Midjourney developed by Midjourney independent research lab, or DALL·E, developed by OpenAI.

[260] *Ibidem.*

[261] J. D. ZAMFIRESCU-PEREIRA, ET AL., *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*, cit.

writing – or, in tech vernacular, engineering – prompts».[262] Prompt engineering has thus emerged as a widely demanded skill within the workplace. Indeed, a new «cottage industry has already sprung up around those who can speak to the machines. On PromptBase, a marketplace for prompt engineers, you can purchase a few lines of text to feed into any number of generative-AI models».[263] In slightly less than two years since the release of ChatGPT in November 2022, mastering the art of 'whispering to the genie' to unravel its algorithmic knowledge[264] has not only become a profitable and beneficial competence for personal and professional development, but has also evolved into a real profession, sought-after and highly remunerated.

While the degree of efficiency can vary depending on the technique employed and the specific LLM used, numerous empirical studies have quantified the performance enhancements achieved by means of this approach. For instance, the chain of thought (CoT) technique, which aims to facilitate coherent and stepwise reasoning processes in LLMs by simulating human-like reasoning applied through intermediate logical steps in problem solving, has been shown to achieve 90.2% accuracy rate in experiments with the gsm8k benchmark – a dataset comprising different mathematical problems – utilising the PaLM 540B model.[265] In another study conducted by the American Academy of Orthopaedic Surgeons (AAOS), the reflection of thoughts (RoT) technique, which

[262] CHARLIE WARZEL, *The Most Important Job Skill of This Century*, in «The Atlantic», 8 Feb. 2023, https://www.theatlantic.com/technology/archive/2023/02/openai-text-models-google-search-engine-bard-chatbot-chatgpt-prompt-writing/672991/, accessed on 20.07.2024.

[262] *AI Act*, in *European Commission, Digital Strategy*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, accessed on 20.07.2024.

[263] *Ibidem*.

[264] ARAS BOZKURT, RAMESH C. SHARMA, *Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World*, in «Asian Journal of Distance Education», vol. 18, n. 2, 2023, https://doi.org/10.5281/zenodo.8174941, accessed on 20.05.2024.

[265] JASON WEI, ET AL., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, in *ArXiv*, 10 Jan. 2023, https://doi.org/10.48550/arXiv.2201.11903, accessed on 29.04.2024. PaLM (Pathways Language Model) is a transformer architecture-based LLM developed by Google AI and announced in April 2022.

encourages the LLM to review and reflect on its previous steps by simulating a discussion between experienced individuals, showed the highest consistency with medical guidelines on osteoarthritis management. This technique achieved an overall consistency rate of 62.9% and a 77.5% consistency rate for strong recommendations, using GPT-4.[266] These examples illustrate how well-designed prompts can significantly enhance the coherence and reliability of model responses, thereby fostering more successful AI-human interactions.

In the educational domain, prompt engineering eases the co-creation of learning materials specifically tailored to the individual needs of students. This targeted approach enhances teaching assistance efficacy, making the educational process more engaging and personalised. A study conducted on the potential of prompt engineering to improve the teaching of computer programming, revealed how it facilitates learning, complex problem-solving, and improves the performance of student-generated code. Notably, the multi-step conversational strategy, which comprises dynamic interactions between the LLM and the student with iterative feedback and refinement, demonstrated a 100% success rate with the utilisation of the GPT-4 model.[267] Teachers can also use LLMs to create interactive quizzes that dynamically adapt to the student's level of knowledge,[268] providing immediate feedback and thus enabling increased learning awareness and efficiency. Not only does the use of clearly defined prompts leverage LLMs to support

[266] LI WANG, ET AL., *Prompt engineering in consistency and reliability with the evidence-based guideline for LLM*s, in «Npj Digital Medicine», vol. 7, n.41, 2024, https://doi.org/10.1038/s41746-024-01029-4, accessed on 20.07.2024.

[267] TIANYU WANG, NIANJUN ZHOU, ZHIXIONG CHENG, *Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation*, in *ArXiv*, 07 Jul. 2024, https://doi.org/10.48550/arXiv.2407.05437, accessed on 20.07.2024.

[268] YOSHIJA WALTER, *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*, in «International Journal of Educational Technology in Higher Education», vol. 21, n. 15, 2024, https://doi.org/10.1186/s41239-024-00448-3, accessed on 21.07.2024.

teachers and educators, it also allows students to interact more effectively with these systems, maximising learning outcomes and engagement.[269] Furthermore, prompt engineering can support students in tackling complex tasks by fostering creativity, critical thinking, and self-assessment. They can be encouraged to explore problems from multiple perspectives, developing original solutions, and even critically evaluating the outputs generated by generative AI, thereby gaining a deeper understanding of its inherent limitations and cultivating a more robust AI literacy,[270] which promotes awareness and responsible use.

Despite these advantages, new concepts such as automatic prompt engineer (APE) have been developed and deepened. APE involves utilising LLMs themselves to generate and refine their own prompts rather than relying on human-written ones. This technique considers prompt creation as a black-box optimisation problem, where the model iteratively generates a list of potential prompts, refines them, and selects the best one according to a predefined scoring function. This iterative process continues until an effective solution is reached. Often, this strategy outperforms human-crafted prompts, particularly for complex tasks, demonstrating the potential of LLMs to enhance their own capabilities through self-optimisation.[271] While automated methods such as APE are effective in generating prompts related to general tasks, they may encounter difficulties with highly specialised and complex domains that require in-depth knowledge, such as medical, legal or scientific fields, where technical terminology is crucial. A prompt

[269] A. BOZKURT, R. C. SHARMA, *Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World*, cit.

[270] EMILY THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, paper presented at the 22nd International Conference of the Italian Association for Artificial Intelligence (AIXIA), Rome, Italy, 2023.

[271] YONGCHAO ZHOU, ET AL., *Large Language Models Are Human-Level Prompt Engineers*, in *ArXiv*, 10 Mar. 2023, https://doi.org/10.48550/arXiv.2211.01910, accessed on 03.08.2024.

engineer (understood as someone who is aware and skilled in crafting efficient prompts) is better equipped to capture technical details, contextual nuances, and idiomatic expressions. Furthermore, a human prompt engineer is cognisant of the ethical implications and potential inherent biases reflected in generated outputs. The flexibility and adaptability of a human interlocutor also allows the LLM to rapidly iterate on prompts, responding to changing requirements. This is why, while methods for automating the prompt generation can be helpful, prompt engineering remains a core competence to foster an augmented intelligence paradigm.

The scientific literature has identified several best practices and specific prompt engineering techniques essential for efficient prompt building, which will be discussed in greater detail subsequently. Nevertheless, a well-constructed prompt should possess specific key features (based on the requirements of the task). For instance, the CLEAR framework[272] aims to provide a standard and structured methodology for crafting effective and coherent prompts for LLMs, by emphasising five essential characteristics: Clear, Logical, Explicit, Adaptive, and Reflective.

Clear

«A concise prompt removes superfluous information, allowing AI language models to focus on the most important aspects of the task, resulting in more pertinent and precise responses».[273] If a prompt is clear and specific, the LLM can better comprehend the user's request, thus lessening the possibility of misinterpretation and improving the accuracy of the generated responses. A clear prompt implies a simple, direct and unambiguous

[272] L. S. Lo, *The CLEAR path: A framework for enhancing information literacy through prompt engineering*, cit.
[273] *Ibidem.*

instruction, avoiding vague or intricate language. An example of specificity could be a prompt of the following type:

```
Explain the benefits of using generative AI in education in no
more than three points and their impact on learning.
```

This prompt not only specifies the number of points to be listed (three), but also the expectation of a brief description of the impact of LLMs, guiding the model to provide a detailed and targeted response. The combination of clarity and specificity allows for precise and relevant answers, preventing confusion and the generation of incorrect or inaccurate information.

Logical

Prompts must be structured in such a manner as to follow a coherent and rational sequence, organising information in an order the LLM can easily follow, reflecting the natural flow of human thought or sequential argumentation. «A logically structured prompt enables AI models to better comprehend the context and relationships between various concepts, resulting in more accurate and coherent outputs».[274] In other words, the prompt has to guide the model through a clear and intuitive path of thought, minimising ambiguities and limiting free interpretation that could lead to any confusing answers. For instance:

```
First, provide a brief definition of generative artificial
intelligence (Generative AI). Next, describe the benefits of
```

[274] *Ibidem.*

```
using Generative AI in the creative field, highlighting the
advantages in terms of innovation and automation of the
creative process. Finally, compare these benefits with the
main ethical challenges associated with the use of Generative
AI, such as the generation of fake content and the potential
impact on privacy.
```

In this case, the prompt follows a clear logical sequence: an introduction to the topic, a list of the benefits and advantages in a specific context, and a comparison between the benefits and the main ethical challenges.

Explicit

When crafting a prompt, it is essential to clearly understand and define what is to be achieved by the LLM (in terms of structure, length, and so forth). For instance, a prompt might aim to obtain in-depth information regarding a particular topic, a simple explanation, a comparison between concepts, or a practical example. Specifying the objective of a prompt may also include the definition of the structural format in which the response is to be obtained (template pattern):[275] a table, a list, a paragraph, or a question-answer style response. This instruction directly influences how the output is structured and the presentation of the content:

```
List three benefits of reinforcement learning in a bulleted
list format, including a short description for each point.[276]
```

[275] J. White, et al., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.

[276] Sabit Ekin, *Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices*, in *Authorea TechRxiv*, 04 May 2023, https://www.techrxiv.org/doi/full/10.36227/techrxiv.22683919.v1, accessed on 21.07.2024. Interestingly, ChatGPT was listed as a co-author of the paper.

In this case, the model is not only driven in the content but also in the architecture of the output. Also the inclusion of contextual details and practical examples can facilitate the understanding of the query to structure the output accordingly.[277] For instance, information including the specific domain or field of application, or preliminary and background information, could be included:

```
Explain the concept of reinforcement learning using the
example of a dog being trained by receiving a biscuit as a
reward every time it correctly carries out a command.
```

Context, however, can also take the form of a scenario specifying the setting or situation in which the model should act: this is particularly useful in tasks such as creative writing or dialogue simulation. For instance, in the persona pattern, the prompt is formulated as follows:

```
You are a machine learning university professor: explain the
ethical implications of generative AI to a group of freshmen.
```

In this case, the scenario context helps to define the tone of voice, role and perspective that the LLM takes into consideration for the response. When the context concerns a specific domain, such as medicine, law or engineering,[278] it is beneficial to include technical terms and references to key concepts from that field:

```
Describe the architecture of a convolutional neural network
```

[277] *Ibidem.*

[278] L. WANG, ET AL., *Prompt engineering in consistency and reliability with the evidence-based guideline for LLM*s, cit.

> (CNN) used for image recognition, explaining how the various layers of the network (such as convolution, pooling and connected layers) contribute to the image classification process.

It might be advantageous in some cases to provide a progressive context, adding additional information (e.g. use cases, specific details) as the conversation progresses, gradually constructing it. To summarise, a well-defined context helps the model to interpret prompts correctly, focusing on specific aspects of the problem and providing more precise answers. Creating prompts with proper context requires a deep understanding both of the topic and the expected purpose of the model, as well as the ability to foresee how the information provided will affect the model's output.[279] «Good prompts aren't just specific. They seem to reflect a deeper understanding of the model you are trying to manipulate».[280]

Adaptive

The adaptive approach allows the prompt to be iteratively refined based on the results obtained from the generative AI model, thus progressively improving the quality of the responses generated by the model. Prompts are systematically modified in various formats and contents to determine which configuration produces the most accurate and relevant results. It is an approach that harkens back to iterative prototyping typical of HCI, in which several solutions are tried out before determining the optimal one.[281]

[279] A. BOZKURT, R. C. SHARMA, *Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World*, cit.

[280] C. WARZEL, *The Most Important Job Skill of This Century*, cit.

[281] J. D. ZAMFIRESCU-PEREIRA, ET AL., *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*, cit.

Reflective

An essential part of the iteration process is the evaluation of the effectiveness of the modified prompt. This can be done at an individual level, by testing the new prompt in a specific conversational context to see if it elicits improved and more correct responses, and then at a broader level, by implementing it in a wider range of contexts. Continuous evaluation and iterative modification of prompts allows for constant optimisation of interactions with LLMs.

In addition to these technical and operational indications, the development of safe and ethical prompts for LLMs is essential to ensure their utilisation is both appropriate and does not cause harm: this involves avoiding formulations that may lead the model to generate violent, discriminatory or misleading content. For instance, it is critical to avoid prompts on sensitive topics without providing clear context and direction to responsibly guide the output. A further crucial aspect concerns the design of prompts that encourage neutral responses or the inclusion of different perspectives to avoid the reinforcement of existing stereotypes inherited from the model during the pre-training phase. [282] Additionally, prompts should be employed so that privacy and dignity of individuals are respected. It is advisable, also in view of security threats, to avoid the inclusion of sensitive personal data.[283]

Moving from general definitions to practice, there are several actual prompt engineering techniques and patterns which constitute a crucial aspect of optimising interactions with LLMs. These approaches are diverse and have evolved significantly in

[282] Y. WALTER, *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*, cit.

[283] *Ibidem.*

recent years, thanks to extensive empirical studies that have refined efficient methods to maximise the quality of the output generated. This thesis will consider only the main and most useful ones for academic environments, which have been specifically the subject of the literacy course delivered to students in the empirical phase of the study.

Zero-shot or input-output (IOP)

The large-scale training of LLMs enables them to perform certain tasks in zero-shot (or input-output)[284] mode, which is the technique typically employed in LLMs interactions, as it is the most intuitive. It consists of formulating a query in a simple format, without including instructions, examples, or context: instead, the model relies solely on the knowledge acquired during training (thanks to generalisation capabilities) and the interpretation of the prompt provided.

`Tell me about large language models.`[285]

It is useful when generic answers are needed or when the context is not particularly complex: for instance, it can be used to ask for basic information, or to perform simple tasks such as text classification or proofreading. The zero-shot approach proves to be limited in all those scenarios where more detailed or context-specific answers are required: since it relies on a basic input, it does not allow the model to develop a deeper comprehension, to tailor the answer or to perform complex inferences.[286]

[284] J. WEI, ET AL., *Finetuned language models are zero-shot learners*, in *ArXiv*, 08 Feb. 2022, https://doi.org/10.48550/arXiv.2109.01652, accessed on 23.04.2024

[285] P. LIU, ET AL., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods*, cit.

[286] *Ibidem.*

Few-shot

Despite the significant zero-shot capabilities LLMs acquire due to their sheer size, it is a limiting technique and can lead to vague and unspecific outputs. The few-shot technique takes it a step further: in fact, it implies the provision of a limited number of examples (usually up to five) within the prompt to guide the model's output on specific tasks (one-shot technique also exists, where a single example is provided).[287] This strategy takes advantage of in-context learning, which enables the model to be refined on more specific tasks without further training. Indeed, the few-shot is particularly useful when a limited training dataset is available or when a rapid adaptation of the model to a new task is desired without the need for a lengthy fine-tuning process or large annotated datasets, making LLMs more accessible and usable in various application scenarios, e.g. for automatic question generation (AQC) in educational contexts such as language teaching.[288] To provide another example, the few-shot approach is useful if the model is to generate textual content in a specific language style: for instance, when writing a haiku, one or more poems can be provided as examples, allowing the model to learn the 'pattern', tone of voice or context to be replicated in its future responses.

```
Write an haiku on generative AI. Here are some examples of
haiku:
haiku 1: "I write, erase, rewrite / Erase again, and then / A
poppy blooms"
haiku 2: "The light of a candle / Is transferred to another
candle – / Spring twilight"
```

[287] *Few-Shot Prompting*, in *Prompt Engineering Guide*, https://www.promptingguide.ai/techniques/fewshot, accessed on 13.04.2024.

[288] UNGGI LEE, ET AL., *Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education*, in «Education and Information Technologies», vol. 29, 2024, https://doi.org/10.1007/s10639-023-12249-8, accessed on 1.05.2024.

haiku 3: “The taste / Of rain / – Why kneel?”[289]

While the few-shot is proven to be effective in improving the quality of LLMs outputs, it also has some limitations. In particular, its dependence on the provided example data: if the examples are not representative or are ambiguous, the LLM may generate inaccurate results or misinterpret the context. Furthermore, the effectiveness of few-shot prompting varies significantly depending on the complexity of the task and the quality of the examples: in more complex cases, it may be useful to break the problem into intermediate steps and demonstrate this to the model (chain of thought). Also, «Sometimes, a well-crafted zero-shot prompt can be more effective than providing multiple examples».[290]

Template pattern

The template pattern technique seeks to ensure that the output of an LLM respects a precise, well-defined format and structure that would not normally be employed for the specific type of content being generated.[291] «For example, the user might need to generate a URL that inserts generated information into specific positions within the URL path».[292] An example of usage for this technique for university students might be the generation of emails expressing job interest to various companies:

```
I am going to provide a template for your output. Everything
in brackets is a placeholder. Any time you generate text, try
to fit it into one of the placeholders that I list. Please
```

[289] Taken from THEA VOUTIRITSAS, *10 Vivid Haikus to Leave you Breathless*, in *Read Poetry*, https://www.readpoetry.com/10-vivid-haikus-to-leave-you-breathless/, accessed on 21.08.2024.

[290] As cited in B. CHEN, ET AL., *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*, cit.

[291] J. WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.

[292] *Ibidem.*

```
preserve the formatting and overall template that I provide.
Template:
Dear [company_name],
I am writing to express my sincere interest in the
[internship_position] internship at your company. [...]
I look forward to the possibility of working at [company_name].
[...]
```

The first statement informs the LLM it has to follow a specific format or structure for its output. The second instructs it on using placeholders, guiding where to replace the information obtained in subsequent prompts. The third statement serves to constrain the LLM so that it does not modify the template provided.[293] In subsequent interactions, it is sufficient to write:

```
Generate a motivation letter for an internship_position:
ethical AI at company_name: OpenAI.
```

However, it is important to bear in mind that a consequence of this technique is filtering of the output generated, possibly eliminating other useful information for the user.[294]

Persona pattern, role-play or expert prompting

It is a technique entailing the assignment of a specific role or ‘persona’ to an LLM for emulating the behaviour and characteristics (especially in terms of tone of voice and language style) of a given agent or character. It relies on the idea that these models can generate more relevant and accurate responses if they target the cognitive and behavioural

[293] *Ibidem.*

[294] *Ibidem.*

skills of a specific role model, such as a university professor. «[LLMs] can convincingly mimic various personas, ranging from fictional characters to historical and contemporary figures. The assigned role provides context about the LLM's identity and background. By adopting the persona, the LLM can generate more natural, in-character responses tailored to that role».[295]

```
You are a machine learning university professor: explain the
ethical implications of generative AI to a group of freshmen.
```

This capability has numerous practical applications: for instance, the LLM-powered chatbots developed by Character.AI[296] allow users to create and interact with customised chatbots, not only such as psychologists, study tutors and professors, but also historical figures such as Leonardo da Vinci, Cleopatra, Vincent Van Gogh, or well-known contemporary characters such as Elon Musk.[297] However, role-play prompting can also be useful in multi-agent collaborative environments, through the assignment of specific roles, each acting as a specialised 'expert' in a phase of the work process (e.g. chief executive officer, product manager, auditor, and so on), an approach known as 'expert prompting'.[298] Several studies have demonstrated the effectiveness of the persona pattern in increasing the relevance, consistency and accuracy of the responses generated.[299] Despite the advantages, though, this approach does present some challenges, such as the need to construct complex prompts and the risk of generating erroneous responses when

[295] Aobo Kong, et al., *Better Zero-Shot Reasoning with Role-Play Prompting*, in *ArXiv*, 15 Mar. 2024, https://doi.org/10.48550/arXiv.2308.07702, accessed on 24.05.2024.

[296] https://www.character.ai/

[297] While also raising further ethical and legal issues.

[298] Yu-Min Tseng, et al., *Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization*, in *ArXiv*, 26 Jun. 2024, https://doi.org/10.48550/arXiv.2406.01171, accessed on 19.08.2024.

[299] *Ibidem.*

the AI model does not correctly interpret the assigned persona. In addition, users have to be aware they should not enter personal or sensitive details regarding existing individuals, due to the privacy and security implications of the prompts.

Chain-of-thought (CoT) and zero-shot chain-of-thought (zero-shot CoT)

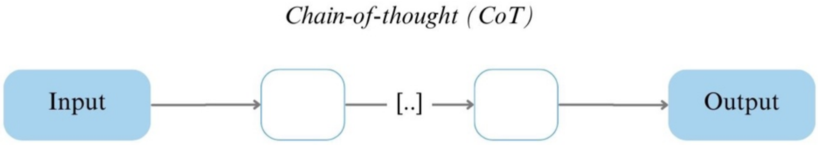


FIGURE 12, *Chain-of-thought (CoT).*

The CoT technique is an advanced approach employed to improve the reasoning capabilities of LLMs. It consists in providing intermediate reasoning steps within the prompt's instructions to steer the model through a logical sequence of steps to solve a complex problem.[300] This technique not only results in more accurate outputs, but also improves the transparency of the model, allowing it to follow an already determined reasoning process.[301] CoT has brought significant achievements in improving the performance of LLMs in advanced problems, such as solving sophisticated mathematical problems over various benchmarks.[302] Its effectiveness stems mainly from its ability to break down intricate problems into smaller, more manageable steps, simulating a human thinking process, and allowing the LLM to deal with each of these steps separately before arriving at the final solution.

[300] J. WEI, ET AL., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, cit.
[301] P. LIU, ET AL., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, cit.
[302] *Ibidem.*

```
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis
balls. Each can has 3 tennis balls. How many balls does he
have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each
is 6 tennis balls. 5 + 6 = 11. The answer is 11.
Q: The cafeteria had 23 apples. If they used 20 to make lunch
and bought 6 more, how many apples do they have?[303]
```

By having a definite reasoning to follow (the proposed example), the model will be able to generate a coherent response. An empirical study further explored why CoT is so successful and, specifically, which aspects of the reasoning demonstration steps contribute to its performance. Results showed how prompts with invalid reasoning demonstrations could still achieve a high percentage of CoT performance, suggesting that the correct order of the steps and their relevance to the query are more influential than the logical consistency of the individual reasoning steps.[304] The CoT findings have crucial implications for the future development of models of LLMs, strengthening their reasoning capabilities. However, it is a complex technique in the prompt conception itself, as it implies the human user has understood in first instance the steps to be performed in order to achieve a solution of the problem. Furthermore, the quality of CoT responses is highly dependent on the model's ability to understand the context, and on the quality of the prompt at the outset. If it is ambiguous or badly phrased, the LLM may provide an incorrect or overly complicated explanation.[305]

[303] Example taken from J. Wei, et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, cit.

[304] Boshi Wang, et al., *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters*, in «Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics», vol. 1, pp. 2717-2739, https://aclanthology.org/2023.acl-long.153, accessed on 04.07.2024.

[305] J. Wei, et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, cit.

While traditional CoT involves providing examples where the individual steps of the reasoning process are made explicit, in contrast, zero-shot CoT simplifies this approach by eliminating the need to provide examples to the LLM. Instead, it employs a simple augmentation of the prompt, by simply adding at the end

```
Let's think step by step.
```
[306]

This modification encourages the model to independently adopt a methodical approach to problem solving, producing a chain-of-thought that is more likely to result in accurate outputs. Typically, this approach comprises two steps. At first, the LLM is prompted to generate a sequential reasoning process through a sentence such as 'Let's think step by step'. Subsequently, a second prompt is used to extract the final answer from the generated chain of thought (if not already answered).[307] Zero-shot CoT has shown significant improvements over the traditional zero-shot technique, especially in contexts where creating examples for a CoT is impractical or unfeasible. One of its crucial limitations is the reliance on the model's inherent capabilities, insufficient in fields where more specialised knowledge is required. Nevertheless, its ease of application and the successful results obtained do make it among the most widely utilised prompt engineering techniques.

[306] TAKESHI KOJIMA, ET AL., *Large Language Models are Zero-Shot Reasoners*, in ArXiv, 29 Jan. 2023, https://doi.org/10.48550/arXiv.2205.11916, accessed on 19.03.2024.
[307] *Ibidem.*

Context manager pattern

The context manager technique in prompt engineering provides a structured method to adjust the setting or context information influencing the model's output, increasing the accuracy, relevance and consistency of the generated responses.[308] As aforementioned, context is a crucial component in the development of a prompt, as it defines the circumstances and conditions under which an LLM has to operate, and it has significant impacts on the output. Specifically, «[t]he Context Manager pattern aims to emphasize or remove specific aspects of the context to maintain relevance and coherence in the conversation».[309] Statements to be included within the prompt regarding what to consider or ignore are, for instance::

```
Within scope X.
Please consider Y.
Please ignore Z.[310]
```

Where X, Y and Z should list key concepts, facts or instructions. The more explicit these statements are, the more the LLM is likely to answer consistently.

```
When analyzing the following pieces of code, do not consider
formatting or naming conventions.[311]
```

In many situations, it might be useful to ask the model to ignore all the previously provided information within the conversation, requesting it to start again, resetting the

[308] J. WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.

[309] *Ibidem.*

[310] *Ibidem.*

[311] *Ibidem.*

context built. In many situations, it might be useful to ask the model to ignore all the previously provided information within the conversation, requesting it to start again, resetting the context built. As in the case of the persona pattern, there is a risk of unintentionally discarding any potentially useful information the LLM would have included and of which the user is unaware. A solution could be to ask for an explanation of which topics would be lost before proceeding.[312]

Flipped interaction pattern

The flipped interaction pattern aims at improving the user interactions with LLMs by reversing the traditional interaction flow, driving the model to generate questions instead of answers.[313] «The goal [...] is to flip the interaction flow so the LLM asks the user questions to achieve some desired goal».[314] Such an approach is particularly effective in the educational field, strengthening the already existing pedagogical paradigm of 'classroom-flipping', as in the case of peer instruction. «Peer Instruction [...] encourages students to share their understandings with peers who in turn challenge their interpretation during group activities».[315] Specifically, through this prompt engineering technique, it is possible to create interactive quizzes, creating student-centred questions and facilitating self-regulated learning through the LLMs' ability to provide immediate feedback, which can be utilised either by teachers or autonomously to adjust learning strategies. During the Covid-19 pandemic, these approaches enabled active student involvement in distance learning modalities. A suggested structure for the prompt is:

[312] *Ibidem.*

[313] CHEE WEI TAN, *Large Language Model-Driven Classroom Flipping: Empowering Student-Centric Peer Questioning with Flipped Interaction*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.14708, accessed on 15.06.2024.

[314] J. WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.

[315] *Ibidem.*

A) I would like you to ask me questions to achieve X

B) You should ask questions until this condition is met or to achieve this goal (alternatively, forever)

C) (Optional) ask me the questions one at a time, two at a time etc.[316]

From now on, I would like you to ask me questions to deploy a Python application to AWS. When you have enough information to deploy the application, create a Python script to automate the deployment.[317]

If a precise number or format for questions (such as multiple choice ones) is not specified, they will be semi-random. The prompt can then be adjusted as required, but the more specific it is, the better the result will be. In addition, when developing a flipped interaction with LLMs, it is necessary to consider the user's level of knowledge and involvement: whether the intention is to achieve the objective with as little user interaction as possible (minimum control), or the opposite (maximum control), it has to be explicitly stated.[318] The flipped interaction pattern favours a greater interactive involvement and deeper understanding on the part of the users, while also stimulating reasoning, self-analysis and critical thinking.

[316] J. WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.

[317] *Ibidem*.

[318] *Ibidem.*

Self-consistency

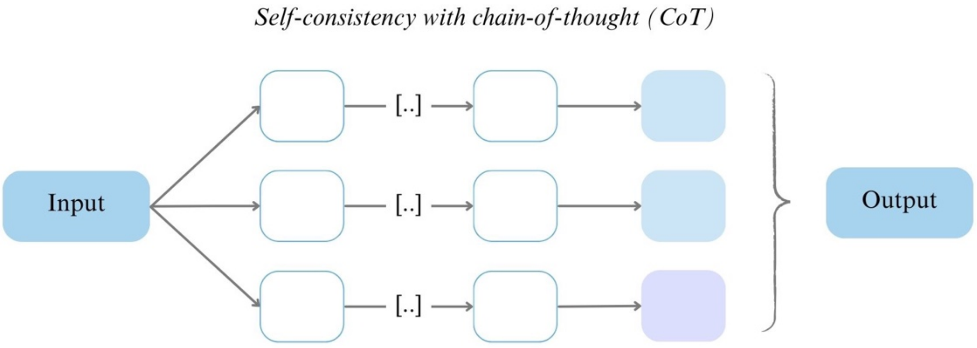


FIGURE 13, *Self-consistency with chain-of-thought (CoT).*

It is one of the most advanced prompt engineering approaches, introduced «to replace the naive greedy decoding used in chain-of-thought prompting»,[319] the decoding technique LLMs employ to select, at each step, the token with the highest likelihood of being the next one in the textual sequence (section I.2), given the context already generated. As the model selects the most likely token (although more advanced LLMs, such as GPT, now employ techniques such as top-p sampling), repetitive or predictable outputs tends to be generated. Notably, less likely alternative paths that could lead to more creative or useful answers are not explored. Through the use of self-consistency, conversely, a series of distinct reasoning paths (traditional CoTs) are provided to the model, then choosing the output most suitable for solving the problem. Thus, in practice, the exact same CoT prompt is supplied iteratively, a number of times at the user's discretion. In the case of mathematical problems, for instance, the output appearing to be the most consistent with the others, i.e. the one that is repeated or most frequently included within the outputs, should be chosen.[320] The combination of CoT and self-

[319] XUEZHI WANG, ET AL., *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, 07 Mar. 2023, https://doi.org/10.48550/arXiv.2203.11171, accessed on 09.05.2024.
[320] *Ibidem.*

consistency produced significant enhancements, resulting, for example, in a 17.9% increase over the gsm8k benchmark already used to test CoT.[321]

Alternative approaches pattern

The alternative approaches approach refers to the employment of LLMs for prompting them to suggest alternative ways of solving a task or problem. The reason is that «[h]umans often suffer from cognitive biases that lead them to choose a particular approach to solve a problem even when it is not the right or "best" approach. Moreover, humans may be unaware of alternative approaches to what they have used in the past».[322] For instance, in the academic environment, students can leverage this technique to investigate different research questions, develop innovative solutions to a problem, suggest various case studies to address, and so on. The structure of an alternative approaches prompt should be:

```
A) Within scope X, if there are alternative ways to accomplish
   the same thing, list the best alternate approaches
B) (Optional) compare/contrast the pros and cons of each
   approach
C) (Optional) include the original way that I asked
D) (Optional) prompt me for which approach I would like to
   use.
```
[323]

```
Whenever I ask you to deploy an application to a specific cloud
service, if there are alternative services to accomplish the
same thing with the same cloud service provider, list the best
```

[321] *Ibidem.*
[322] J. WHITE, ET AL., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, cit.
[323] *Ibidem.*

```
alternative services and then compare/contrast the pros and
cons of each approach with respect to cost, availability, and
maintenance effort and include the original way that I asked.
Then ask me which approach I would like to proceed with.[324]
```

The versatility of this technique allows it to be suitable for a wide range of tasks and scenarios, from technical problem solving to strategic consulting. Furthermore, it encourages a more informed, aware and critical decision-making process, fostering user reflection without delegating responsibility for the decision to the LLM.

Reflection pattern

As detailed in Chapter I.2, LLMs are essentially ‘black boxes’ whose reasoning process leading to an output is difficult to detect, yet with some prompt engineering techniques such as the reflection pattern, the model can be prompted to explain, analyse and evaluate the quality and correctness of the responses it produces.[325] «This pattern is particularly effective for the exploration of topics that can be confused with other topics or that may have nuanced interpretations and where knowing the precise interpretation that the LLM used is important»,[326] such as within academic environments.

```
Whenever you generate an answer
Explain the reasoning and assumptions behind your answer
(Optional) ...so that I can improve my question.[327]
```

[324] *Ibidem*.

[325] *Ibidem*.

[326] *Ibidem*.

[327] *Ibidem*.

A crucial aspect is the ability of this technique to generate internal feedback within the LLM, not only stimulating it to critically re-evaluate the output provided, but also to consider possible alternatives or areas for improvement. Nevertheless, a limitation arises when users do not understand the topic being discussed, as in the case of highly specialised or technical answers, resulting in a cumbersome explanation they cannot comprehend.

Tree-of-thoughts (ToT)

The tree-of-thoughts (ToT) technique exploits a tree-of-thought framework to improve LLMs' ability to solve complex problems. Specifically, a the tree-of-thought is a structure in which every node represents a thinking or semantic unit for solving a task (i.e., coherent linguistic sequences functioning as intermediate steps in problem solving), and the arcs between the nodes denote the logical dependencies between these steps.[328]

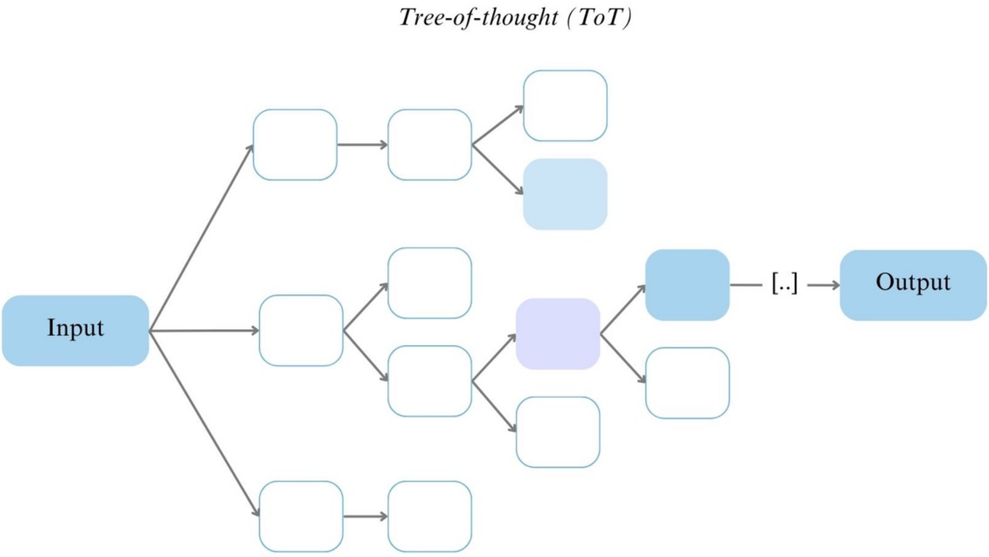


FIGURE 14, Tree-of-thoughts (ToT).

[328] SHUNYU YAO, ET AL., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, in *ArXiv*, 03 Dec. 2023, https://doi.org/10.48550/arXiv.2305.10601, accessed on 05.04.2024.

For instance, in a discussion between experts in a given field, each presents their arguments for a step-by-step resolution of the problem. If one of the pursued pathways reveals errors or proves ineffective, they return to the previous step to explore another pathway. This process proceeds until all the experts involved in the discussion agree on the best solution. Specifically, ToT in prompt engineering is based on the combination of established approaches, namely CoT and self-consistency: it extends these techniques by allowing the LLM to generate a tree-of-thoughts, self-evaluating and reviewing multiple logical paths, going back when inconsistencies are detected, and refining the process until an optimal response is achieved.[329]

```
Imagine three different experts are answering this question.
All experts will write down 1 step of their thinking,
then share it with the group.
Then all experts will go on to the next step, etc.
If any expert realises they’re wrong at any point then they
leave.
The question is... [330]
```

And then the specific problem to be solved is provided within another prompt. Thus, the ability of these models to both generate and evaluate thoughts is combined with search algorithms «which allow systematic exploration of the tree of thoughts with lookahead and backtracking».[331] In this case, the LLM will adopt a proposed breadth-first search (BFS) algorithm, systematically examining all the nodes within a level of the tree before moving on to the next one. The BFS will ensure all possible actions for solving the

[329] *Ibidem.*

[330] Dave Hulbert, *Using Tree-of-Thought Prompting to boost ChatGPT's reasoning*, in «GitHub repository», 2023, https://doi.org/10.5281/zenodo.10323452, accessed on 20.04.2024.

[331] *Ibidem.*

problem will be evaluated before choosing the one to lead to the next level. A disadvantage occurs when the tree-of-thoughts has a high branching, as the BFS becomes particularly complex and may exhaust the output window of the LLM.[332] The ToT technique has been shown to significantly improve the performance of LLMs on a number of complex tasks, such as the Game of 24, a mathematical game in which the player is required to use four numbers and the basic arithmetic operations (addition, subtraction, multiplication and division) to obtain the result of 24, where they were 74% successful compared to 9% with traditional CoT prompting. [333] Thanks to its capacity to flexibly and thoroughly navigate complex scenarios, selecting the most efficient pathways, ToT allows the LLM to be more robust against hallucinations. This approach is also particularly useful in educational scenarios and advanced problem-solving, where it is relevant to consider multiple factors or assumptions. In creative writing, for instance, it can be employed for generating and evaluating different storylines or narrative developments, simulating an iterative brainstorming process. It is currently an evolving technique: future research could explore node search algorithms such as Monte Carlo Tree Search (MCTS), which combines random simulation techniques (Monte Carlo) with a tree-of-thought framework to explore possible moves or choices.[334]

[332] S. YAO, ET AL., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, cit.
[333] *Ibidem.*
[334] *Ibidem.*

FOURTH CHAPTER

ARTIFICIAL INTELLIGENCE LITERACY (AIL)

"At the end of the day, an algorithm is just a recipe. If we do not understand this, and continue to imagine Carl Sagan's aliens or Stanley Kubrick's sentient computers, we will not be able to develop the cultural antibodies we will need to coexist with our creatures."
NELLO CRISTIANINI, *La scorciatoia*

"Ipsa scientia potestas est.
Knowledge is power."
SIR FRANCIS BACON, *Meditationes Sacrae*

IV.1. Navigating the future with awareness

«What kinds of capabilities do people need in a world infused with AI? How can we conceptualise these capabilities? How can we help learners develop them? How can we empirically study and assess their development?».[335] The revolution in terms of intelligent systems capable to achieve complex goals, the access to knowledge, and in terms of human-AI interactions, is raising numerous concerns about the capabilities people need to consciously navigate in this rapidly changing and fundamentally altered

[335] L. MARKAUSKAITE, ET AL., *Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?,* cit.

reality. Just as «the appearance of computers in the workplaces at the turn of the 21st century has added 'algorithmic thinking' and 'computing literacy' to the repertoire of thinking skills and literacies that have been seen as essential for successful functioning and employment in society»,[336] so too with the advancement of GAI and a re-distribution of intelligence, decision-making and labour, there has emerged a demand for new competencies to collaborate effectively and ethically with such technologies. Despite the countless different definitions the scientific community has attempted to attribute to this concept, some common foundations can be identified. «AI literacy means having the essential abilities that people need to live, learn and work in our digital world through AI-driven technologies».[337] It can be defined as «an individual's ability to clearly explain how AI technologies work and impact society, as well as to use them in an ethical and responsible manner and to effectively communicate and collaborate with them in any setting. It focuses on knowing (i.e. knowledge and skills)».[338] In summary, it is a comprehensive and multifaceted framework that not only enables users to critically and collaboratively utilise AI and GAI systems in everyday life, education and work, but also to understand how they operate and critically assess their inherent limitations and ethical and legal implications. This involves a set of critical competences embracing both knowledge and expertise, along with critical evaluation, ongoing self-reflection and a continuous learning attitude required to keep up with such rapid and abrupt changes. Furthermore, increasingly, frameworks are being explored in which competencies are not

[336] *Ibidem.*

[337] DAVY TSZ KIT NG, ET AL., *Conceptualizing AI literacy: An exploratory review*, in «Computers and Education: Artificial Intelligence», vol. 2, 2021, https://doi.org/10.1016/j.caeai.2021.100041, accessed on 02.05.2024.

[338] DURI LONG, BRIAN MAGERKO, *What is AI Literacy? Competencies and Design Considerations*, in «CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems», pp. 1-16, https://doi.org/10.1145/3313831.3376727, accessed on 20.04.2024.

just individual but also collective and interconnected, integrating ecological and network dynamics principles (i.e. models considering the interaction between various systems and the environment), and are highly dependent upon the social and cultural context in which individuals operate.[339] AIL is crucial for individuals to be able to knowingly and effectively navigate in a society where these systems play a significant role in decision-making processes at every level and constitute invaluable helpers in many daily personal and professional tasks, as misuse or improper design could cause (even irreparable) personal and collective damage. «Even though AI literacy is regarded as a future skill, studies that examine how it may affect a user's behavior when dealing with LLM-based AI systems like ChatGPT are currently lacking»[340] and «to the best of our knowledge, a closer look at the AI literacy of individual target groups through literature analysis is still lacking».[341] As a matter of fact, while research literature has extensively examined and discussed the meaning boundaries of artificial intelligence literacy and the abilities which would be indispensable to fulfil this purpose, there is still little empirical evidence and concretisation of these intentions. It is precisely in this grey area where this experimental thesis aims to place itself.

A skill shared by the majority of AI literacy research to date is the understanding of the fundamental concepts relating to artificial intelligence: what it is, how it functions, the basic underlying principles (such as machine learning, neural networks, and the stochastic prediction of successive tokens within a textual sequence in the case of

[339] MICHAL ČERNÝ, *University Students' Conceptualisation of AI Literacy: Theory and Empirical Evidence*, in «Social Sciences», vol. 13, n. 3, 2024, https://doi.org/10.3390/socsci13030129, accessed on 29.08.2024.

[340] NILS KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, in «Computers and Education: Artificial Intelligence», vol. 6, 2024, https://doi.org/10.1016/j.caeai.2024.100225, accessed on 20.08.2024.

[341] MATTHIAS CARL LAUPICHLER, ET AL., *Artificial intelligence literacy in higher and adult education: A scoping literature review*, in «Computers and Education: Artificial Intelligence», vol. 3, 2022, https://doi.org/10.1016/j.caeai.2022.100101, accessed on 03.03.2024.

LLMs),[342] together with the «features that make an entity 'intelligent', including discussing differences between human, animal, and machine intelligence».[343] Specifically in this regard, a suitable definition of AI clearly setting out the functionalities and purposes of these technologies is the one provided by the European Commission in the *Ethics Guidelines for Trustworthy AI* previously mentioned:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.[344]

Comprehending the teleological purpose behind the definition of intelligence, therefore taking into consideration different forms of intelligence beyond the anthropocentric one that usually tends to be considered, is crucial for dehumanising and streamlining these systems, while at the same time recognising their pragmatic usefulness in several domains.[345] Furthermore, it is equally paramount for users to be aware of the inherent limitations deriving from AI and GAI models' architecture and functioning, such as hallucinations, limited comprehension capability, lack of semantics, inability to reach a thorough level of logical and mathematical reasoning, as well as limited capability to maintain context.[346] «Understanding the current capabilities of AI – and that there are still

[342] D. TSZ KIT NG, ET AL., *Conceptualizing AI literacy: An exploratory review*, cit.

[343] D. LONG, B. MAGERKO, *What is AI Literacy? Competencies and Design Considerations*, cit.

[344] EUROPEAN COMMISSION, *Ethics Guidelines for Trustworthy AI*, cit.

[345] N. CRISTIANINI, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano,* cit.

[346] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

many open questions in AI research [...] – can help users in making more informed decisions».[347] For with such a ground-breaking and widespread rise of ChatGPT and generally GAI, the fallacious opinion that has spread widely is «the misleading perception that LLMs can effortlessly provide solutions across domains»,[348] as if they were 'oracles' directly feeding into the whole knowledge base of mankind and the world: one consequence of the ELIZA effect and automation bias discussed in chapter II. Critically confronting the inherent limitations of AI not only enables it to be utilised with greater awareness, exploiting its strengths to the full, but allows at the same time an understanding of the importance of human-oversight and validation within the co-creation process. It is crucial to trigger in citizens a critical and well-structured consideration of how their assessment and reasoning skills must be integrated into the interaction stream with artificial intelligence systems, in order to avoid a depletion of decision-making faculties, prevent misaligned or potentially harmful outputs for oneself or one's neighbour, and to foster collaborative human-AI teaming interactions.

A further fundamental competence constituting AI literacy is the awareness and knowledge of the ethical and legal implications involved in the functioning of these systems, such as privacy, accountability, fairness, transparency (addressed specifically in section I.4), notwithstanding that the interpretation and prioritisation of these principles may vary significantly depending on the cultural, social and institutional context.[349] Particularly, it is important for individuals to be aware of the potential risks and challenges arising from the introduction of these technologies into society. «There are many ethical questions surrounding how AI should be used, and there has been growing

[347] D. Long, B. Magerko, *What is AI Literacy? Competencies and Design Considerations*, cit.

[348] E. Theophilou, et al., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit

[349] Arif Ali Khan, et al., *Ethics of AI: A Systematic Literature Review of Principles and Challenges*, in *ArXiv*, https://doi.org/10.48550/arXiv.2109.07906, accessed on 13.09.2024.

concern surrounding issues such as AI's effect on the job market, bias and discrimination in AI, and AI-related data privacy scandals».[350] Ethical awareness in AI literacy implies not only the ability to critically evaluate how these technologies affect society, but also to reflect on the consequences of the generated outputs. This is an essential requirement for ensuring their responsible use and that their applications are designed with human welfare and social justice in mind. Moreover, with the emergence of strict regulatory frameworks such as the AI Act, it is imperative for citizens to be empowered to recognise behaviours, systems, trends that are or are not legally compliant. Both in the educational context, where educators and trainers should integrate these themes in their curricula to prepare students in becoming informed and responsible citizens in the age of GAI, and in the work context, where trust needs to be maintained between the organisation and its stakeholders and workers' rights need to be protected, this kind of AI literacy is of paramount importance. Nevertheless, despite the great acknowledgement this set of skills and knowledge has proven to gain, there is a significant deficiency in AI ethics literacy programmes within academic institutions.

Several studies have shown how although many educational entities have begun to incorporate these notions into their programmes, these efforts are often fragmented and lack a standardised approach.[351] For instance, UNESCO developed an in-depth framework for AI ethics that was adopted globally by all member states in 2021. Among the guidelines whose adoption it promotes there are awareness and literacy. Specifically, «[p]ublic understanding of AI and data should be promoted through open & accessible education, civic engagement, digital skills & AI ethics training, media & information

[350] D. LONG, B. MAGERKO, *What is AI Literacy? Competencies and Design Considerations*, cit.

[351] ANDREA ALER TUBELLA, MARÇAL MORA-CANTALLOPS, JUAN CARLOS NIEVES, *How to teach responsible AI in Higher Education: challenges and opportunities*, in «Ethics and Information Technology», vol. 26, n. 3, 2023, https://doi.org/10.1007/s10676-023-09733-7, accessed on 08.08.2024.

literacy»,[352] also developing a guidance for the use of AI in education and research. Even in the business context, AI ethics literacy will become crucial for professionals working in digital work environments. «Employees need to become informed stakeholders about the future of work, and provided with opportunities to develop their foundational knowledge and skills. Only then could they engage with future endeavors of AI design and use it as AI-empowered workers».[353]

While understanding the functioning at a high level of the stochastic and algorithmic processes that lie at the foundation of AI is essential, as well as the implications of these systems being integrated in real-world scenarios, programming skills or prior knowledge of computer science is not. «“AI literacy” encompasses AI competencies that the general population should possess and accordingly focuses mainly on learners without a computer science background (“non-experts”)».[354] Precisely in light of the shifts in terms of increased accessibility of human-AI conversational interactions discussed in chapter II, it can be argued that «[t]he hottest new programming language is English».[355] For non-technical users, it is much more relevant to understand how to interact with GAI systems not through computer code language, rather through natural language optimised for interaction with machines. «Higher-quality prompt engineering skills predict the quality of LLM output, suggesting that prompt engineering is indeed a required skill for the goal-directed use of generative AI tools».[356] «[...] educational

[352] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris, UNESCO, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en, accessed on 15.04.2024.

[353] DILEK CETINDAMAR, ET AL., *Explicating AI Literacy of Employees at Digital Workplaces*, in «IEEE Transactions of Engineering Management», http://dx.doi.org/10.1109/TEM.2021.3138503, accessed on 05.09.2024.

[354] M. C. LAUPICHLER, ET AL., *Artificial intelligence literacy in higher and adult education: A scoping literature review*, cit.

[355] ANDREJ KARPATHY, «The hottest new programming language is English», in *X*, https://x.com/karpathy/status/1617979122625712128?lang=en, accessed on 02.04.2024.

[356] N. KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, cit..
E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

promptization can arguably be considered to be part of a future AI literacy. An essential component of it is that one has to encompass the relational and symmetrical engagement with AI models».[357] Thus, to round off the knowledge and skills set forming AIL's basis, there is also pragmatic training in the formulation of effective prompts. Despite the accessibility and efforts to smooth and humanise GAI performances, being able to interrogate technology correctly, and also knowing how to critically evaluate the generated outputs in order to iteratively refine them, represents a promising and challenging frontier. «However, research on the perspectives of non-experts using LLM-based AI systems through prompt engineering and on how AI literacy affects prompting behavior is lacking».[358]

The actual problem facing educators, and the wider society, is the fact that technological progress is advancing at a much faster pace than conceptual, ethical, normative, and educational advancement. This instance will necessarily give rise to a reconsideration of the ways and terms in which learning takes place, encouraging a comprehensive and diversified literacy, and ultimately fostering an attitude of continuous and personalised learning to coexist and cooperate with artificial intelligence.

[357] HADVAN HAUGBAKEN, MARIANNE HAGELIA, *A New AI Literacy For The Algorithmic Age: Prompt Engineering Or Educational Promptization?*, in «2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)», 2024, http://dx.doi.org/10.1109/ICAPAI61893.2024.10541229, accessed on 15.06.2024.

[358] N. KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, cit..

IV.2. Education

The sheer impact AI and GAI is permeating numerous sectors, including education, from primary level education to university and adult education. The AIED concept involves the implementation of these technologies within educational practices at different levels. It extends further than the development of personalised teaching and learning systems, such as virtual tutors (ITS), or automated evaluation tools. the aim of AIED is to maximise the effectiveness of education through the use of advanced tools enhancing the experience and acquisition of knowledge and skills of any type of 'learner', understood in the broadest sense of the term. For instance, ChatGPT Edu, is an educational version of the tool designed to provide a teaching and learning facilitator to assist and ease learning, improve understanding of complex topics, and deepen students' engagement with knowledge. «We built ChatGPT Edu because we saw the success universities like the University of Oxford, Wharton School of the University of Pennsylvania, University of Texas at Austin, Arizona State University, and Columbia University in the City of New York were having with ChatGPT Enterprise»,[359] i.e. the version dedicated to companies. At the European level, the EU developed in 2022 the *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*,[360] which are part of the Digital Education Action Plan (DEAP, 2021-2027), comprising a series of policy initiatives in order to create a common standard within the member states with regard to high quality, inclusive and accessible digital

[359] *Introducing ChatGPT Edu*, in *OpenAI*, 20 May 2024, https://openai.com/index/introducing-chatgpt-edu/, accessed on 02.09.2024.

[360] EUROPEAN COMMISSION, *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*, Bruxelles, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2766/153756, accessed on 13.08.2024.

education. Specifically, the approach adopted by the EU takes into account both the potential of these systems in personalising learning to better meet the individual needs of students, or in promoting a more flexible and responsive teaching approach, and the ethical challenges in ensuring they are fair, free of bias, transparent and explainable.[361] However, in parallel to implementing artificial intelligence within the educational context, disseminating appropriate literacy becomes a crucial aspect accordingly. Within the context of DEAP, the EU Commission has recognised that strong competence in the use of technologies such as AI is crucial for the development of a sustainable digital economy. AI literacy not only prepares students to interact with complex systems, but also supports ethical and responsible use, reducing the risks of inappropriate or unethical use of data.

There are notable educational initiatives attempting to encourage the implementation of AIL within schools, such as AI4K12, sponsored by the Association for the Advancement of Artificial Intelligence (AAAI) and the Computer Science Teachers Association (CSTA), and aimed at developing both guidelines and resources for teaching AI from kindergarten through high school (K-12).[362] A central point of the initiative is the introduction of the five key principles establishing the major dimensions which should be included in AIL: perception, representation and reasoning, learning, natural interaction and social impact. With regard to the context of higher and adult education, several governments have recognised the importance of artificial intelligence literacy. For instance, the Federal Ministry of Education and Research (BMBF) in Germany has published a directive to encourage it, with a double objective. Firstly, to

[361] ID., *AI report – By the European Digital Education Hub's Squad on artificial intelligence in education*, Bruxelles, Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2797/828281, accessed on 13.08.2024.

[362] https://ai4k12.org/, accessed on 10.08.2024.

broaden the availability of specialised academic workforce for economic and scientific development within the field of AI, and secondly, to promote the utilisation of AI to improve the quality of higher education.[363] On the other hand, the Finnish government funded the creation of *Elements of AI*,[364] one of the earliest free and easily accessible online courses jointly developed by the University of Helsinki in cooperation with the technology company Reaktor, and launched in 2018 with the aim of providing an introduction to the fundamentals of AI to a wide audience without a technical background. More than one million students from 170 countries worldwide have participated in the course to date. Furthermore, France, Italy, Slovenia, Ireland and Luxembourg co-developed AI4T (artificial intelligence for and by teachers),[365] funded by Erasmus+, an initiative enabling teachers to be trained on artificial intelligence, through innovative learning methods such as various MOOCs (Massive Open Online Courses) and open textbooks.

With regard to the Italian context specifically, through the strategic plan *Strategia Italiana per l'intelligenza artificiale* (*Italian Strategy for Artificial Intelligence)*,[366] AI literacy is intended to be broadened by providing structured upskilling and reskilling programmes within companies in various sectors and public administration, as well as by implementing AI learning paths in schools, creating apprenticeships, and introducing this research field into several university degrees, in addition to supporting the National PhD in AI (which besides covering cutting-edge aspects of technological and scientific

[363] FEDERAL MINISTRY OF EDUCATION AND RESEARCH (BMBF), *Richtlinie zur Bund-Länder-Initiative zur Förderung der Künstlichen Intelligenz in der Hochschulbildung*, 2021, retrieved from https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2021/02/3409_bekanntmachung.html, accessed on 02.09.2024.

[364] https://www.elementsofai.com/, accessed on 03.04.2024.

[365] https://www.ai4t.eu/, accessed on 16.08.2024.

[366] *Strategia italiana per l'intelligenza artificiale 2024-2026*, cit.

development, has a dedicated curriculum for investigating the ethical and social implications of these technologies). It is important to highlight how, according to this strategy, AIL should transcend any field of university degree study, regardless of whether or not students have a technical background. «[...] one of the risks associated with the rapid development of techniques and knowledge in the field of AI lies in the limited access to continuous and up-to-date training opportunities for users, which would enable them to build the necessary skills and abilities to understand its costs and benefits, critically assess its processes, and creatively use its tools. To mitigate this risk, it will therefore be essential to integrate foundational teachings on Artificial Intelligence into all university courses, including non-STEM subjects, with content tailored to the objectives of the specific disciplines».[367] EDUNEXT – Next Education Italy, funded with 22.4 million euro by the Ministry of University and Research, and involving 35 universities and 5 higher education institutions for art and music, is also intended to innovate the approach to university learning by focusing on strategic competences, including AI, and by promoting inclusive and accessible digital training paths.[368] Several Italian universities offer national MOOCs on AI. The Carlo Bo University of Urbino covers the topic from a scientific, technical, ethical, social, economic, philosophical and cultural perspective, without any prerequisite requirements for participants,[369] and the Federico II University of Naples provides a four-track programme ranging from the basics of AI to machine learning and introduction to programming,[370] amongst others. Furthermore, there are several Bachelor's, Master's and post-graduate courses in AI offered by Italian

[367] Translated from *Ibidem*.

[368] PRESS OFFICE UNIMORE, *EDUNEXT. Al via il progetto per l'innovazione della formazione digitale a livello nazionale*, University of Modena and Reggio Emilia, 2024, https://www.magazine.unimore.it/site/home/notizie/articolo820069931.html, accessed on 09.09.2024.

[369] https://mooc.uniurb.it/wp/aimooc/

[370] https://www.federica.eu/federica-pro/intelligenza-artificiale-e-scienza-dei-dati/

universities, for instance the University of Bologna's international Master's degree programme in *Artificial Intelligence*, the University of Pisa's Master's degree programmes in *Artificial Intelligence and Data Engineering* and in *Biotechnologies and Applied Artificial Intelligence for Health*, the University of Milan-Bicocca's and University of Pavia's Master's programme in *Human-centred Artificial Intelligence*, to mention but a few. As far as cross-curricular training for students not belonging to STEM degree courses is concerned, Ca' Foscari University of Venice has recently reorganised the syllabus of the Computer Skills course for the academic year 2025/2025, targeting language and humanities students, now entirely focused on artificial intelligence and practical applications within these fields.[371]

The Italian educational landscape is evolving, yet attempts are still fragmentary and individualised, lacking a collective framework for collaboration and the establishment of standards to support the integration of AI literacy in the various curricula. This approach is likely to generate inequalities in access to AI competences, leading to a discrepancy between more advanced institutions and those less prepared for the digital revolution. Greater cooperation and accessibility of educational resources, also for alumni and citizenship in general, would be advantageous. AI literacy should become a transversal objective extending beyond the usual learning spaces and modalities, in order to educate and inspire an aware citizenry capable of interacting with these systems for good. Another critical aspect is the absence of empirical data demonstrating the effectiveness of current policies and programmes. The integration of pragmatic studies, highlighting successes, failures, or areas for improvement, could help to critically assess them, providing the

[371] *Abilità informatiche: nuovo corso sull'IA per lauree umanistiche e lingue*, in *Ca' Foscari University of Venice*, 03 Sep. 2024, https://unive.it/pag/14024/?tx_news_pi1%5Bnews%5D=15770&cHash=4eebcb25d37fc994bb7619b7308c4c55, accessed on 05.09.2024.

research community with a more comprehensive perspective and the possibility to deepen the scope of study for further improvement in the future. Educational institutions, governments and international organisations should collaborate not only to establish guidelines, but also to create continuous monitoring and evaluation mechanisms allowing for real-time adjustments and the adaptation of training courses to rapid changes in the field. Lastly, although knowledge acquisition regarding the functioning and definition of AI systems is crucial, the absence of a robust ethical component in AIL could lead to irresponsible use of AI and GAI, with long-term negative consequences both socially and culturally. Therefore, these pathways should firmly take into account the importance of user empowerment in terms of the ethical and legal, human-in-the-loop and fallacy implications of these technologies.

SECOND PART

FIFTH CHAPTER

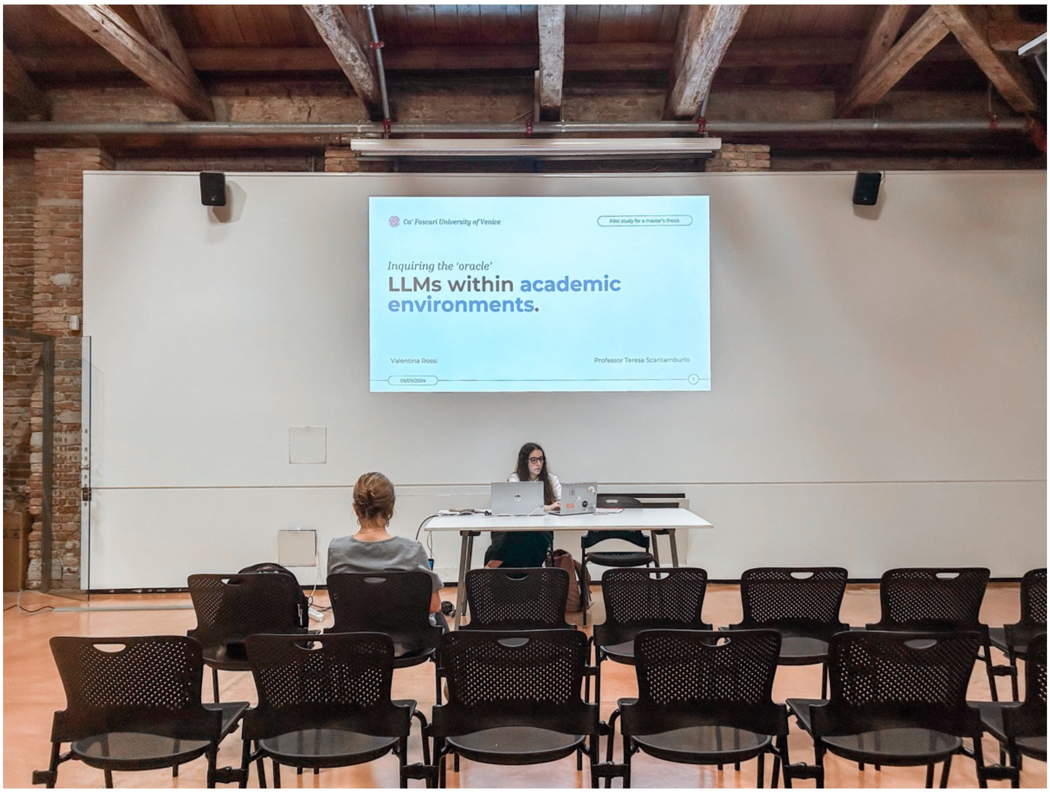
CONTEXT AND SCOPE OF THE EMPIRICAL STUDY

> *"ChatGPT is incredibly limited but good enough at some things to create a misleading impression of greatness. It's a mistake to be relying on it for anything important right now. it's a preview of progress; we have lots of work to do on robustness and truthfulness."*
>
> SAM ALTMAN, CEO of OpenAI, X 12/10/2022

V.1. Purposes and expected results

The primary purpose of the pilot study is to evaluate whether a literacy course on generative artificial intelligence (GAI) and prompt engineering can improve the understanding, awareness, and interaction skills of non-technical university students with LLMs-based systems, with a particular focus on ChatGPT, due to its considerable popularity to date. The project's objectives encompass the improvement of students' abilities to formulate effective prompts, the further development of their knowledge and opinion regarding GAI and LLMs, as well as the increase of ethical awareness and critical reflection about the utilisation of such technologies not only within academic environments, but also in everyday and professional contexts. This study represents an innovative approach to AIL in higher education, as it focuses on targeting non-STEM students. It is a fundamental aspect, since in a context where AI is transforming both labour dynamics and personal lives, restricting literacy to technical experts alone risks excluding a significant part of the population from critical discussions about technology and the skills needed to consciously interact with it. Nevertheless, redesigning AIL for non-technical profiles requires a pedagogical approach going beyond mere technical literacy. What is needed is a didactic approach relating the functioning of AI and GAI tools with their concrete applications and social and ethical implications. To this end, a learning framework was established integrating theoretical and practical modules, fostering the acquisition of both knowledge and skills through the experimentation of prompt engineering techniques and prompt patterns, which made explicit the risks and opportunities offered by GAI models, in particular LLMs. Another innovative aspect of this study is its empirical nature. Despite a growing interest in literacy, the majority of studies focus on formulating theoretical definition and adoption frameworks, as discussed in chapter IV.

The study was implemented through an attendance workshop (Figure 15) lasting approximately three and a half hours, preceded and followed by two questionnaires and practical exercises with ChatGPT. It is hypothesised that the proposed educational initiative will produce a measurable positive impact in the knowledge, awareness and interaction skills of the participants, confirming the importance of educating users from non-STEM backgrounds in the effective and aware utilisation of AI and GAI at university and beyond. Specifically, there are three research questions.



Ca' Foscari University of Venice

Pilot study for a master's thesis

Inquiring the 'oracle'

LLMs within academic environments.

Valentina Rossi

Professor Teresa Scantamburlo

06/09/2024

1

FIGURE 15*, Initial slide presented during the workshop.*

What is the effect of literacy on students in terms of formulating effective prompts?

Prompt engineering represents a new and crucial competence for effective interaction with GAI, through the conversational interfaces with which it is engineered. Defined as the ability to structure prompts (the inputs provided to the LLM) in a precise way to elicit high quality responses from AI models, this is a competence developed with practice and an understanding of the systems' inner functioning and intrinsic limitations. Several studies have highlighted how inexperienced users tend to use ChatGPT or other LLMs with an unsystematic approach, often dealing with these tools as 'humans' rather than as machines with specific constraints, and with excessive expectations regarding the kind of intelligence that would be expected from them, actually far from transcendental omniscience.[372] It is hypothesised that AI literacy, specifically technological knowledge of how these systems work but especially a structured and in-depth introduction to the techniques of prompt engineering and prompt patterns, will foster a positive impact on students' prompt engineering skills, consequently increasing the quality of prompts for specific tasks.[373] Furthermore, by teaching students how to formulate structured prompts, it encourages them to critically reflect on the information they are seeking and the context in which they operate, thus helping them to articulate their needs more clearly, encouraging deeper understanding and self-analytical skills.[374] It can be assumed, therefore, that the workshop will improve the formulation of effective and aware prompts with the LLM, and a more critical thinking attitude.

What is the effect of literacy on students in terms of their opinion and knowledge of GAI and LLMs?

Literacy regarding AI, GAI and in particular LLMs allows increasing the knowledge and understanding of non-technical students regarding these technologies, enabling them to distinguish between AI and non-AI systems, identify use cases, and understand the role these models play in human-AI interaction.[375] «An important target for AI literacy, involving LLM, is defusing the rising and misleading feeling of being able to access and process any form of knowledge to solve problems in any domain with no effort or previous expertise in AI or problem domain».[376] Therefore, it is hypothesised

[372] N. KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, cit.
[373] D. J. WOO, ET AL., *Effects of a Prompt Engineering Intervention on Undergraduate Students' AI Self-Efficacy, AI Knowledge, and Prompt Engineering Ability: A Mixed Methods Study*, cit.
[374] Y. WALTER, *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*, cit.
[375] N. KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, cit.
E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.
[376] *Ibidem.*

that the workshop will enable students to understand the main concepts related to this field, critically evaluating AI technologies and taking more account of the inherent limitations during their interactions and validation of the generated outputs. Furthermore, it is supposed that literacy may not only provide knowledge, but also the ability to develop more nuanced and less extremist opinions towards the threats posed by AI. Literate students might see GAI and LLMs not as completely positive or completely negative technologies, but as tools that can be used in both constructive and challenging ways: conversely, lack of literacy might lead to more drastic views, such as irrational enthusiasm for AI or exaggerated fear.[377] Accordingly, AIL would not only provide students without a technical background with a greater knowledge of how GAI operates, but also of its inherent limitations, enabling them to be aware of these during their interactions with the models, while at the same time rationalising the use of these tools in real-world contexts, as helpful assistants requiring, nevertheless, careful human supervision.

How does literacy contribute to student's ethical awareness of the use of GAI?

The third hypothesis concerns the impact of AI literacy on ethical awareness with regard to the use of LLMs and GAI more generally. Previous studies indicate that, when structured training in this area is provided, students become more aware of the limitations and ethical implications of these technologies, leading to a more grounded understanding of the risks of delegating decision-making faculties exclusively to these systems, without validation and critical analysis.[378] Literacy plays a crucial role as it empowers students to exercise critical thinking in the interpretation and use of AI-generated knowledge,

[377] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

[378] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

encouraging the maintenance of the central role of human control in interactions with these systems, specifically in light of issues such as plagiarism, data privacy and the potential for misinformation.[379] Furthermore, AI literacy is positively correlated with students' ability to use these technologies in ways that enhance their learning experience, while maintaining ethical considerations, such as monitoring generated content in terms of biases or avoiding over-reliance on AI-generated content.[380] Hence, it is assumed that AIL will enable students to use the tools effectively, but also improve their ability to deal with ethical challenges, making critical thinking a key component of their responsible use.

In summary, this pilot study provides an initial insight into the impact of AIL specifically focused on GAI and prompt engineering on a heterogeneous sample of humanities students: the diversity of the participants contributed to delineating a range of experiences and perceptions, demonstrating how these students, although often less familiar with advanced technological tools, can benefit from targeted training. This first experiment offers a solid basis for developing future large-scale studies, with the aim of exploring the influence of these initiatives not only in the academic sphere, but also in the students' transversal and personal skills, making them more aware, balanced, and critical of the use and consideration of these technologies.

[379] JINHEE KIM, ET AL., *Exploring students' perspectives on Generative AI-assisted academic writing*, in «Education and Information Technologies», 2024, https://doi.org/10.1007/s10639-024-12878-7, accessed on 20.08.2024.

[380] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

C. ZHAI, S. WIBOWO, L. D. LI, *The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review*, cit.

V.2. OpenAI

This sub-chapter will address the history of OpenAI, the founding company behind ChatGPT, with a special focus on the evolution of its LLMs. Silicon Valley is a northern California area that, *nomen omen*, initially owed its fortune to a robust industrial network dedicated to the production of semiconductors and microchips (hence the reference to silicon), and it has since become the catalyst and metonymy for American, and, by extension, Western technological development. In 2015, its dynamism was palpable, not only concentrated around colossal industries dominating the sector (such as Apple Inc., Google Inc.,[381] Facebook, Inc.,[382] Intel Corporation, Cisco Systems, Inc. and Nvidia Corporation, which established their headquarters here); but also witnessing the emergence of numerous start-ups.[383] It is precisely in this atmosphere of fervent innovativeness that OpenAI was founded in December 2015[384] as a «non-profit artificial intelligence research company»,[385] with an initial capital of USD 1 billion. Intrinsically connected to this economic, scientific, and, first and foremost, socially relevant framework since its establishment, its founders and funders include names that were already widely known and engaged. Ilya Sutskever, research scientist at the Google Brain Team, Greg Brockman, former Stripe, Inc. CTO, «[t]he group's other founding members are world-class research engineers and scientists: Trevor Blackwell, Vicki Cheung, Andrej Karpathy, Durk Kingma, John Schulman, Pamela Vagata, and Wojciech

[381] Now Google LLC. See *Google*, in *Wikipedia*, https://en.wikipedia.org/wiki/Google, accessed on 14.02.2024.

[382] Now Meta Platforms, Inc. See *Meta Platforms*, in *Wikipedia*, https://en.wikipedia.org/wiki/Meta_Platforms, accessed on 14.02.2024.

[383] BRIAN SOLOMON, *The Hottest Startups Of 2015*, in «Forbes», 17 Dec. 2015, https://www.forbes.com/pictures/eimh45ehmdj/hottest-startups/, accessed on 14.02.2024.

[384] *OpenAI*, in *Wikipedia*, https://en.wikipedia.org/wiki/OpenAI, accessed on 14.02.2024.

[385] GREG BROCKMAN, ILYA SUTSKEVER, *Introducing OpenAI*, in *OpenAI*, https://openai.com/blog/introducing-openai, accessed on 14.02.2024.

Zaremba».[386] Peter Andreas Thiel, PayPal Holdings, Inc. co-founder and former CEO, Samuel Harris Altman, Y Combinator president, and Elon Reeve Musk, CEO of Tesla, Inc. acted as co-chairs.

Within the declaration of intent published on the company's official website, it is stated:

> Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact.
>
> We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as possible. The outcome of this venture is uncertain and the work is difficult, but we believe the goal and the structure are right. We hope this is what matters most to the best in the field.[387]

The objectives, therefore, were deliberately ambitious, and – as mentioned in the following few lines – given the recent accomplishments of deep neural networks, were aimed at preparing humankind towards the moment of technological singularity[388] (which is portrayed as a necessary and unavoidable stage in the evolutionary development of the human civilisation, although not predicting its advent chronologically). The underlining desire was to support the highest quest of artificial intelligence research ever since the 1956 Dartmouth workshop that heralded its birth: to create an artificial general intelligence (AGI) capable of achieving human levels performance across the wide spectrum of cognition and action. Nevertheless, these premises were followed by

[386] *Ibidem.*

[387] *Ibidem.*

[388] A hypothetical chronological future point at which technological progress surpasses human cognition and the ability to forecast its consequences.

contrasting responses. The real driving force behind machine learning is data, and there were concerns regarding how the openness promised (to the extent of building the company's own name on it) could be combined with sufficient availability to spark progress.[389]

On 27th April 2016, the public beta version of OpenAI Gym was released, i.e. «a toolkit for developing and comparing reinforcement learning (RL) algorithms. It consists of a growing suite of environments (from simulated robots to Atari games), and a site for comparing and reproducing results».[390] On 5th December of the same year, it was instead released «Universe, a software platform for measuring and training an AI's general intelligence across the world's supply of games, websites and other applications».[391] In the following years, the company's expenditures and efforts were focused on strengthening its functional resources, such as cloud computing, and implementing new efficient strategies to train more in-depth and powerful models, all the while with the declared goal of building a general artificial intelligence. It was in June 2018 that a seminal paper was issued destined to change the company's course definitively, namely *Improving Language Understanding by Generative Pre-Training*:[392]

> We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabelled text, followed by discriminative fine-tuning on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve

[389] NEIL LAWRENCE, *OpenAI won't benefit humanity without data-sharing*, in «The Guardian», 14 Dec. 2015, https://www.theguardian.com/media-network/2015/dec/14/openai-benefit-humanity-data-sharing-elon-musk-peter-thiel, accessed on 15.02.2024.

[390] G. BROCKMAN, *OpenAI Gym Beta*, in *OpenAI*, https://openai.com/research/openai-gym-beta, accessed on 16.02.2024.

[391] *Universe*, in *OpenAI*, https://openai.com/index/universe/, accessed on 06.09.2024.

[392] ALEC RADFORD, ET AL., *Improving Language Understanding by Generative Pre-Training*, in OpenAI, 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, accessed on 04.05.2024.

> effective transfer while requiring minimal changes to the model architecture.[393]

Precisely through this publication, in fact, the first language model developed by OpenAI was introduced, namely GPT-1. It was based on the transformer architecture, which had been introduced the year before; however, what made it so innovative was the technique employed for training, differing significantly from the ones previously adopted. As a matter of fact, the language models hitherto developed heavily relied on supervised learning approaches, i.e. extremely large, manually annotated datasets which made development extremely costly and labour-intensive, and also essentially limited to languages with a significant corpora of textual resources available. GPT-1, conversely, was trained through a novel method combining a first phase of unsupervised pre-training with a second phase of supervised fine-tuning. «The closest line of work to ours involves pre-training a neural network using a language modelling objective and then fine-tuning it on a target task with supervision».[394] The first phase, performed by using BookCorpus, a dataset of more than 7000 unpublished books of various genres, enabled the model to learn the generalities and properties of language, without specific labels to guide it in classifying the elements. The second phase, fine-tuning, was designed to adapt the pre-trained parameters to specific tasks using a smaller, labelled dataset, allowing it to perform effectively in a wide range of linguistic challenges, such as question answering and textual entailment. In addition, it introduced the employment of an autoregressive architecture, where the generation of each token within a textual sequence is conditioned on the previous tokens. This approach enabled the model to generate coherent text, one token at a time, by iteratively feeding its previous outputs as inputs for future predictions.

[393] *Ibidem*.
[394] *Ibidem*.

GPT-1's impact has been substantial within the research community, promoting a significant research and development wave, as it demonstrated significant improvements over previous models. Not only did it consolidate the effectiveness of the transformer architecture, but also demonstrated how pre-trained models could be effectively customised with minimal task-specific data.

The following year, in 2019, OpenAI announced its reorganisation from a non-profit to a 'limited profit' company. Any investment whose return exceeded 100 times to would henceforth be transferred to a non-profit organisation, which would then allocate them as deemed appropriate. «We'll need to invest billions of dollars in upcoming years into large-scale cloud compute, attracting and retaining talented people, and building AI supercomputers. We want to increase our ability to raise capital while still serving our mission, and no pre-existing legal structure we know of strikes the right balance. Our solution is to create OpenAI LP as a hybrid of a for-profit and non-profit – which we are calling a "capped-profit" company».[395] This change raised a lot of criticism: «[t]he new structure has OpenAI LP doing the actual work the company is known for: doing interesting and perhaps widely applicable AI research, occasionally withheld in order to save the world».[396] «Meanwhile, other researchers have bemoaned OpenAI's hyperbole and questioned whether its switch to for-profit research undermines its claims to be "democratizing" AI»,[397] however, it seemed an overriding requirement to be able to compete with the largest tech companies. As previously discussed, developing an AI

[395] *OpenAI LP*, in *OpenAI*, https://openai.com/index/openai-lp/, accessed on 03.09.2024.

[396] DEVIN COLDEWEY, *OpenAI shifts from nonprofit to 'capped-profit' to attract capital*, in «TechCrunch», 11 Mar. 2019, https://techcrunch.com/2019/03/11/openai-shifts-from-nonprofit-to-capped-profit-to-attract-capital/, accessed on 03.09.2024.

[397] JAMES VINCENT, *Microsoft invests $1 billion in OpenAI to pursue holy grail of artificial intelligence*, in «The Verge», 22 Jul. 2019, https://www.theverge.com/2019/7/22/20703578/microsoft-openai-investment-partnership-1-billion-azure-artificial-general-intelligence-agi, accessed on 02.09.2024.

model is extremely expensive, making it extremely difficult for research centres or non-profit companies to achieve ambitious performances. The company was subsequently able to distribute company shares to its employees, and in July 2019 it took a further step by partnering with Microsoft, announcing a billion-dollar investment «to support […] building artificial general intelligence (AGI) with widely distributed economic benefits».[398] Specifically, Microsoft built an Azure-based supercomputer for OpenAI to train its artificial intelligence models, «powered by 285 000 CPU cores and 10 000 GPUs»,[399] which has been in use ever since. In February 2019, nevertheless, a major change occurred among the Silicon Valley research halls. GPT-2 was announced, namely a model whose 1.5 billion parameters represented a significant improvement compared to its predecessor. It was trained on a dataset called WebText, comprising textual data extracted from approximately 45 million website links. Among the most remarkable changes with respect to the traditional transformer architecture discussed in chapter I, was the implementation of an architecture focused exclusively on the decoding part: indeed, the model employs a stack of decoder layers (decoder-only), each one characterised by a masked self-attention mechanism followed by a feed-forward neural network. As opposed to encoder-decoder architectures, where both inputs and outputs are handled, decoder-only models are optimised exclusively for text generation. They allow long text dependencies to be handled, reduce complexity and memory and computation requirements, and at the same time promote more focused training as they concentrate entirely on text generation.[400] Furthermore, important technical changes were introduced

[398] *Microsoft invests in and partners with OpenAI to support us building beneficial AGI*, in *OpenAI*, https://openai.com/index/microsoft-invests-in-and-partners-with-openai/, accessed on 04.09.2024.
[399] *OpenAI*, in *Wikipedia*, cit.
[400] JESSE ROBERTS, *How Powerful are Decoder-Only Transformer Neural Models?*, in *ArXiv*, 02 Feb. 2024, https://doi.org/10.48550/arXiv.2305.17026, accessed on 02.09.2024.

with GPT-2, such as layer normalisation (a method to stabilise the learning process by improving the flow of gradients), moving it to the start of each model sub-block as a residual pre-activation network. The initialisation process of weights was also altered to better take into consideration the depth of the model, by scaling the weights of residual layers at initialisation by a factor inversely proportional to the square root of the number of residual layers themselves.[401] OpenAI released the source code and weights of the GPT-2 model (open-weights), making it publicly accessible on GitHub and allowing developers and researchers to utilise and adapt the model for various research purposes and applications.[402] Performance on various language modelling benchmarks demonstrated improved capabilities and state-of-the-art results, not only technically but also in human interactions. However, GPT-2 was not immediately released due to the emerging capabilities of the model, for which concerns were raised regarding possible risks, especially in terms of misleading content generation and propaganda.[403] «OpenAI's concerns are being taken seriously by some. A team of researchers from the Allen Institute for Artificial Intelligence recently developed a tool to detect "neural fake news"».[404]

These concerns did not daunt OpenAI's research team, which instead maintained a fast pace and in May 2020 released a new language model: GPT-3, which like its predecessor comprises a decoder-only transformer model, pre-trained and then fine-tuned on a more specific dataset. «Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language

[401] A. RADFORD, ET AL., *Language Models are Unsupervised Multitask Learners*, in *OpenAI*, 2019, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, accessed on 03.09.2024.

[402] *Gpt-2*, in *GitHub*, https://github.com/openai/gpt-2, accessed on 02.09.2024.

[403] *GPT-2: 1.5B release*, in *OpenAI*, 5 Nov. 2019, https://openai.com/index/gpt-2-1-5b-release/, accessed on 20.08.2024.

[404] OSCAR SCHWARTZ, *Could 'fake text' be the next global political threat?*, in «The Guardian», 04 Jul. 2019, https://www.theguardian.com/technology/2019/jul/04/ai-fake-text-gpt-2-concerns-false-information, accessed on 04.09.2024.

model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets [...]».[405] Among its innovations, the most significant is the increased amount of training data and, as a result, of parameters. GPT-3 has been trained on a vast dataset called 'Common Crawl', together with other curated datasets in order to ensure both variety and high quality in the training material, such as WebText2, Books1, and the entire English-language Wikipedia encyclopaedia, despite «the entirety of English Wikipedia constitutes just 0.6 percent of GPT-3's training data».[406] It is precisely the enlarged size that enabled the LLM to acquire emergent abilities in new and unfamiliar contexts not previously encountered during training, and to adapt to entirely novel tasks without further fine-tuning. Specifically, in few-shot learning contexts, where the model is exposed to a limited number of examples of a given task prior to being tested on it, due to its ability for in-context learning, GPT-3 has been shown to achieve and outperform previously state-of-the-art models that had been specifically fine-tuned on those tasks. «While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning».[407] As already covered in chapter I, conversely to the gradual improvements to be expected as parameters and training data grow, the emergent abilities are characterised by unexpected and abrupt leaps in performance, which only manifest themselves once the model reaches a certain size scale (i.e. exceeding a certain threshold

[405] T. B. BROWN, ET AL., *Language Models are Few-Shot Learners*, cit.
[406] J. VINCENT, *OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws*, in «The Verge», 30 Jul. 2020, https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential, accessed on 02.09.2024.
[407] *Ibidem*.

of parameters or FLOPs). A common example is LLMs' ability to successfully execute complex arithmetic operations like three-digit addition. In the smallest models (PLMs), this capability does not occur, however, above a certain threshold of sophistication (typically over 10 billion parameters), sudden performance enhancements are observed.[408] The emergent abilities of computer code generation, and creative writing in both prose and poetry, attracted as much attention from the media as from the research community, which also raised numerous ethical concerns. «This month, OpenAI, an artificial-intelligence research lab based in San Francisco, began allowing limited access to a piece of software that is at once amazing, spooky, humbling and more than a little terrifying. OpenAI's new software, called GPT-3, is by far the most powerful "language model" ever created».[409] However, OpenAI CEO Samuel Altman promptly silenced any speculation. «The GPT-3 hype is way too much. It's impressive [...] but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out».[410] Differently from GPT-2, whose weights were made publicly available, OpenAI decided that GPT-3 would only be made accessible through an API, i.e. an interface between different software, enabling them to interact and exchange data or functionalities in a supervised manner without the need to be aware of each other's internal specifications. The GPT-3 API was regulated by a licensing and subscription scheme, implying that users had to pay in order to access the service, and allowed the company to manage its computational

[408] J. WEI, ET AL., *Emergent Abilities of Large Language Models*, cit.

[409] FARHAD MANJOO, *How Do You Know a Human Wrote This?*, in «The New York Times», 29 Jul. 2020, https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html, accessed on 04.09.2024.

[410] S. ALTMAN, «The GPT-3 hype is way too much. It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out», in *X*, https://x.com/sama/status/1284922296348454913, accessed on 02.09.2024.

resources load, optimising model utilisation without overloading the systems. In late 2020 and early 2021, a number of OpenAI employees, including Dario Amodei, formerly Vice President of Research, established Anthropic PBC. A company created out of a strong and avowed commitment to the development of responsible artificial intelligence technologies, prioritising public benefit alongside financial returns, an ethos consolidated by its status as a Public Benefit Corporation.[411]

ImageGPT, announced in June 2020, marked a revolutionary approach in the computer vision domain by employing the transformer. In fact, it was based on the GPT architecture, despite it being originally created for natural language processing, applied here to visual data. The main idea underlying ImageGPT is to predict the next pixel within a sequence of pixels, just as it is possible to predict the next token within a textual sequence.[412] Notably, the core of this model's structure is a transformer decoder-only model, in which images are scaled to a low resolution (32x32, 48x48 or 64x64 pixels) and then converted into a one-dimensional sequence of RGB values, thereby processing the image as a sequence of tokens similar to a text sequence and therefore predicting the value of the forthcoming pixel relying on the previous ones, arranged in a raster sequence (from left to right and from top to bottom). The loss function used for optimising the model is the negative log-likelihood, whereby the model attempts to maximise the logarithmic probability of the predicted pixels in comparison to the actual pixels. This means the prediction is based on maximising the proper chance of the subsequent pixel given the context of the preceding ones. By 2021, thanks to the advancements made with ImageGPT, OpenAI developed a GAI model capable of generating images from textual

[411] SHIKHAR GHOSH, SHWETA BAGAI, *Anthropic: Building Safe AI*, in «Harvard Business School Case 824-129», 2024.

[412] MARK CHEN, ET AL., *Generative Pretraining from Pixels*, in *OpenAI*, https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf, accessed on 19.07.2024.

inputs, relying on a combination of autoregressive transformers and discrete variational autoencoders (dVAE), a component necessary to convert high-resolution images into a compressed representation as discrete tokens: it was called DALL·E.

> DALL·E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs. […] GPT-3 showed that language can be used to instruct a large neural network to perform a variety of text generation tasks. Image GPT showed that the same type of neural network can also be used to generate images with high fidelity. We extend these findings to show that manipulating visual concepts through language is now within reach.[413]

Furthermore, in 2021, OpenAI developed another successful GPT-3-based tool: Codex, an advanced GAI model capable of generating source code from natural language descriptions. Specifically, the LLM was fine-tuned using a vast collection of computer code, approximately 159 GB of filtered data, with the aim of enabling it to generate and debug code in various programming languages.[414] «OpenAI Codex is a descendant of GPT-3; its training data contains both natural language and billions of lines of source code from publicly available sources, including code in public GitHub repositories. OpenAI Codex is most capable in Python, but it is also proficient in over a dozen languages [...]».[415] In comparison with the standard version of GPT-3 previously discussed, changes were made to the standard tokeniser to better handle peculiarities of the computer code, such as significant whitespaces in Python, thereby integrating special

[413] *DALL·E: Creating images from text*, in *OpenAI*, 5 Jan. 2021, https://openai.com/index/dall-e/, accessed on 13.07.2024.

[414] MARK CHEN, ET AL., *Evaluating Large Language Models Trained on Code*, in *ArXiv*, 14 Jul. 2021, https://doi.org/10.48550/arXiv.2107.03374, accessed on 03.09.2024.

[415] WOJCIECH ZAREMBA, GREG BROCKMAN, *OpenAI Codex*, in *OpenAI*, 10 Aug. 2021, https://openai.com/index/openai-codex/, accessed on 03.09.2024.

tokens to represent sequences of spaces.[416] As a result of a partnership with GitHub, a platform allowing developers to create, store and share their code, OpenAI provided this model to enhance it directly within the integrated development environment (IDE), thereby creating GitHub Copilot. The practical application of Codex extends its usefulness from code generation to the automatic correction of bugs, however, there are important security considerations related to code generation, especially when it comes to source code that could be employed in production. Studies raised concerns about its potential to introduce or fail to identify existing vulnerabilities, a crucial issue for trust in software development.[417] In parallel, Codex attracted considerable attention for its potential impact on computer science education, particularly in introductory programming courses, where it proved to outperform many students in typical examination questions.[418] Nevertheless, research indicated that whilst Codex improved code-writing performance, there were concerns that over-reliance could hinder learning and retention, therefore it became apparent how teachers would have to try to balance its benefits against the potential dangers in terms of learning barriers. Furthermore, the Free Software Foundation, a non-profit software freedom organisation, expressed worries regarding a potential copyright infringement by Codex and consequently by GitHub Copilot, raising the sensitive issue of whether training on public repositories falls under ‘fair use’, which as mentioned in chapter I is one of the main pillars of intellectual

[416] *Ibidem.*

[417] HAMMOND PEARCE, ET AL., *Examining Zero-Shot Vulnerability Repair with Large Language Models*, in *ArXiv*, 15 Aug. 2022, https://doi.org/10.48550/arXiv.2112.02125, accessed on 03.09.2024.

[418] JAMES FINNE-ANSLEY, ET AL., *My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises*, in «ACE '23: Proceedings of the 25th Australasian Computing Education Conference», https://doi.org/10.1145/3576123.3576134, accessed on 03.09.2024.

property law in the United States, but also whether deep learning models can themselves be copyrighted and by whom.[419]

On 30th November 2022, OpenAI announced the worldwide release of ChatGPT, an LLM-based chatbot that interacts with users in a conversational manner, which will be discussed in more detail in the following sub-chapter.[420]

V.3. OpenAI's flagship: ChatGPT

«Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback».[421] At the beginning of 2022, OpenAI announced it had developed InstructGPT, a variant of GPT-3 which employed reinforcement learning during the fine-tuning phase in order to align the LLM with human preferences by means of a labelled dataset of demonstrations of the intended behaviour, thus increasing its effectiveness in interacting with users. This technique was called reinforcement learning from human feedback (RLHF), and constituted one of the main assets of ChatGPT.

[419] DONALD ROBERTSON, *FSF-funded call for white papers on philosophical and legal questions around Copilot: Submit before Monday, August 23, 2021*, in *Free Software Foundation*, 28 Jul. 2021, https://www.fsf.org/blogs/licensing/fsf-funded-call-for-white-papers-on-philosophical-and-legal-questions-around-copilot, accessed on 03.09.2024.

[420] *Introducing ChatGPT*, in *OpenAI*, 30 Nov. 2022, https://openai.com/index/chatgpt/, accessed on 11.05.2024.

[421] L. OUYANG, ET AL., *Training language models to follow instructions with human feedback*, cit.

Indeed, thanks to its implementation, hallucinations were reportedly minimised as well as a slight decrease in the generation of toxic outputs, while not compromising the capabilities of the underlying model.[422] The first phase of the RLHF involved the collection of data from human labellers, writing answers to specific prompts sent via OpenAI's API. These responses provided examples of the model's desired behaviour, and were subsequently employed to train it in a supervised context. Secondly, the next step consisted of training a reward model on a new dataset comprising pairs of model outputs from which the human reviewers had to choose the preferred one. The goal was for the reward model to accurately represent human preferences in a quantifiable and actionable manner for the subsequent reinforcement learning phase. During the final stage, utilising the reward model as an actual reward function, the LLM was further trained with the Proximal Policy Optimisation (PPO) algorithm. This process sought to maximise the rewards obtained, driving the system to generate outputs that were increasingly aligned with the human preferences codified in the reward model itself. These phases can be iterated to continue refinement. As the LLM improves, new comparative data can be collected to update and enhance the reward model, which can in turn guide additional reinforcement training. The approach proved to be successful, yet a few months after GPT-4 was released, it emerged how for this fine-tuning phase, OpenAI employed Kenyan workers, underpaid $1.32 to $2 per hour (depending on seniority and performance) through the company Sama, in order to label extremely disturbing and damaging data. «Some of it described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self-harm, and incest».[423] This investigation raised

[422] *Ibidem.*
[423] BILLY PERRIGO, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, in «Time», 18 Jan. 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/, accessed on 07.04.2024.

considerable concerns regarding the staggering labour force underlying the AI industry, often ignored, with workers calling nationally and internationally for greater recognition and involvement to tackle this exploitative condition.

On 30th November 2022, OpenAI released ChatGPT globally:

> We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.[424]

Specifically, ChatGPT was originally built on GPT-3.5, a significant advancement over its predecessor GPT-3 in various technical and functional domains. One of the most notable innovations was the integration of the RLHF training protocol – which had not been employed in previous public versions – whose purpose consisted in enhancing the model's effectiveness in interactive applications such as chatbots and virtual assistants. Additionally, GPT-3.5 introduced optimisations in energy efficiency and the use of computational resources. In terms of robustness, the LLM demonstrated enhanced resilience, particularly in scenarios involving adversarial attacks, such as those related to cybersecurity, as well as in handling out-of-distribution data; despite the persistence of some more deep-rooted vulnerabilities. The most substantial progress, however, was observed in the its performance on specific natural language understanding (NLU) tasks.[425] A major upgrade over GPT-3 was the capability of GPT-3.5 to better cope with code generation assignments, although consistency issues remained with performance

[424] *Introducing ChatGPT*, in *OpenAI*, 30 Nov. 2022, https://openai.com/index/chatgpt/, accessed on 13.03.2024.

[425] *Ibidem*.

fluctuations depending on context and precise prompt formulation (prompt engineering).[426]

ChatGPT immediately became viral, notwithstanding being met with astonishment, excitement or scepticism, to the extent that it reached one million users in approximately one week. It «is, quite simply, the best artificial intelligence chatbot ever released to the general public»,[427] wrote the New York Times closely followed by major international newspapers. As tackled in chapter II, it was its user interface designed to be intuitive and accessible, providing timely and relevant answers without having to write a single line of computer code that determined such widespread popularity. Users could subscribe with their e-mail, and immediately having the possibility to type their requests in natural language directly into a conversational environment under the form of a chat. This constituted a major departure from LLMs usually accessed by highly specialised individuals through computer code. Despite its several advantages, however, it did not take very long for ChatGPT's intrinsic limitations to emerge. GAI systems were discerned as important catalysts and enablers for human intelligence, «[b]ut they do not always tell the truth. Sometimes, they even fail at simple arithmetic. They blend fact with fiction. And as they continue to improve, people could use them to generate and spread untruths»[428] or, for instance, «programming advice platform Stack Overflow temporarily banned answers by the chatbot for a lack of accuracy».[429] Despite these self-evident

[426] JUNJIE YE, ET AL., *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models*, in *ArXiv*, 23 Dec. 2023, https://doi.org/10.48550/arXiv.2303.10420, accessed on 10.09.2024.

[427] KEVIN ROOSE, *The Brilliance and Weirdness of ChatGPT*, cit.

[428] CADE METZ, *The New Chatbots Could Change the World. Can You Trust Them?*, in «The New York Times», 11 Dec. 2022, https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html, accessed on 04.05.2024.

[429] MARCEL SCHARTH, *The ChatGPT chatbot is blowing people away with its writing skills. An expert explains why it's so impressive*, in «The Conversation», 06 Dec. 2022, https://theconversation.com/the-chatgpt-chatbot-is-blowing-people-away-with-its-writing-skills-an-expert-explains-why-its-so-impressive-195908, accessed on 12.05.2024.

constraints, slightly less than two years after its release «OpenAI says that more than 200 million people use ChatGPT each week»,[430] for a wide variety of purposes, ranging from summarisation to the generation of texts for personal, academic or business purposes, to the generation and review of computer code, from information retrieval to assistance in brainstorming and creative processes, and so forth. Its release and such widespread success stimulated the deployment of many competing products, including Meta Platforms' LLaMa on 24th February 2023 and Google LLC's PaLM-E on 10th March 2023.

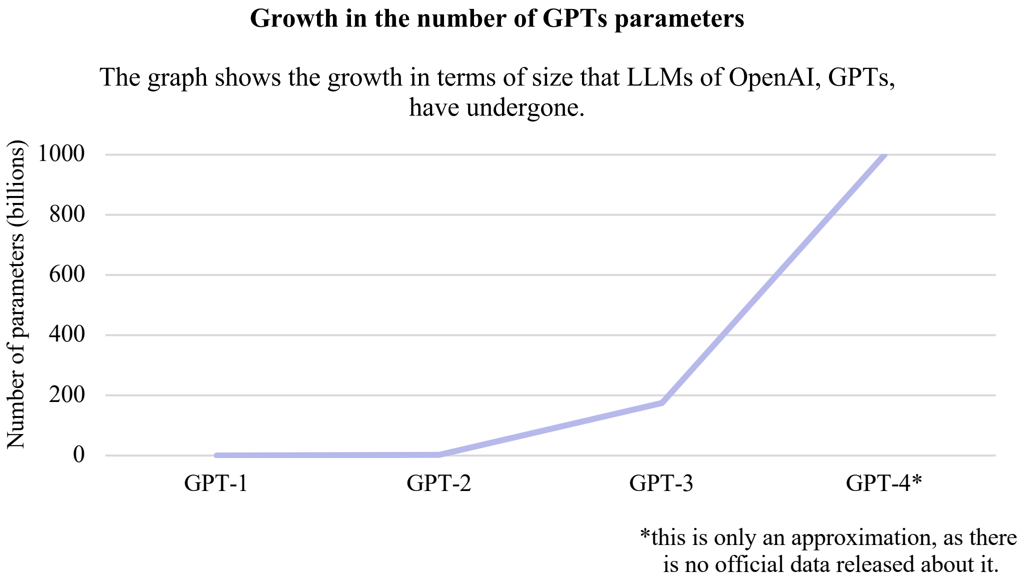
In March 2023, OpenAI released within the chatbot GPT-4, «a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers»,[431] only available through the fee-based version of the tool, i.e. ChatGPT Plus, and through the free Microsoft chatbot Copilot. Although no explicit information was released regarding the precise number, GPT-4 was significantly larger in terms of parameters than its predecessors. It had approximately more than a trillion parameters,[432] leading to increased contextual understanding and reasoning capabilities, especially in multilingual contexts. «Just hours after its release, several users said they created computer games in less than a minute by simply asking the chatbot to generate code, resulting in near-perfect renditions of Tetris, Connect Four, Snake, and Pong. Other users created a matchmaking

[430] INA FRIED, *OpenAI says ChatGPT usage has doubled since last year*, in «Axios», 29 Aug. 2024, https://www.axios.com/2024/08/29/openai-chatgpt-200-million-weekly-active-users, accessed on 02.09.2024.

[431] J. ACHIAM, ET AL., *GPT-4 Technical Report*, cit.

[432] MATTHIAS BASTIAN, *GPT-4 has more than a trillion parameters – Report*, in The Decoder, 25 Mar. 2023, https://the-decoder.com/gpt-4-has-a-trillion-parameters/, accessed on 20.04.2024.

service, bedtime stories, a browser extension that translates any webpage into "pirate speak," and even a tool that can help discover new medications».[433]



*this is only an approximation, as there is no official data released about it.

FIGURE 16, *Growth in the number of GPTs parameters.*

Furthermore, GPT-4 was capable to handle much longer contextual windows than its predecessors, augmenting its ability to maintain coherence in extended conversations or complex documents. Several of these features were forecast before the training of the model, whereas others were found to be difficult to anticipate due to downstream scaling laws breaks.[434] An indicative graph with respect to the evolution of the size of GPTs from GPT-1 to GPT-4 can be seen above (Figure 16). In addition to state-of-the-art capabilities, thanks to the RLHF protocol, it was possible to decrease «the model's tendency to respond to requests for disallowed content by 82% compared to GPT-3.5, and GPT-4 responds to sensitive requests (e.g., medical advice and self-harm) in accordance with

[433] NIK POPLI, *GPT-4 Has Been Out for 1 Day. These New Projects Show Just How Much More Powerful It Is*, in «Time», 15 Mar. 2023, https://time.com/6263475/gpt4-ai-projects/, accessed on 10.09.2024.
[434] J. ACHIAM, ET AL., *GPT-4 Technical Report*

[OpenAI's] policies 29% more often».[435] A couple of days after the release of GPT-4, still in March 2023, OpenAI further introduced support for ChatGPT plugins, including both proprietary ones, such as web browsing and code interpretation, in addition to external ones. As a result, this approach enabled «language models to read information from the internet strictly expand[ing] the amount of content they can discuss, going beyond the training corpus to fresh information from the present day»,[436] and also to run snippets of Python code «in a sandboxed, firewalled execution environment, along with some ephemeral disk space».[437] In November 2023, GPTs were launched, i.e. «a service that allows individuals and small businesses to build customized versions of [...] ChatGPT, and instantly share them on the internet»,[438] through a completely natural language interface, without additional software or the need to develop software code, for instance by providing an additional set of documents specialising in a particular domain. However, the novelties were not restricted to this kind of opportunities. In September 2023, OpenAI announced the release of «a new version of its DALL-E image generator to a small group of testers and folded the technology into ChatGPT»[439] as well as the fact that «ChatGPT can now see, hear, and speak».[440] These advancements were made feasible through the incorporation of different types of data, including text, images, and audio, using advanced, hybrid deep neural networks architectures as detailed in chapter II.

[435] *Ibidem*.
[436] *ChatGPT plugins*, in *OpenAI*, 23 Mar. 2023, https://openai.com/index/chatgpt-plugins/, accessed on 09.09.2024.
[437] *Ibidem.*
[438] CADE METZ, *OpenAI Lets Mom-and-Pop Shops Customize ChatGPT*, in «The New York Times», 06 Nov. 2023, https://www.nytimes.com/2023/11/06/technology/openai-custom-chatgpt.html, accessed on 09.09.2024.
[439] ID., TIFFANY HSU, *ChatGPT Can Now Generate Images, Too*, in «The New York Times», 06 Nov. 2023, https://www.nytimes.com/2023/09/20/technology/chatgpt-dalle3-images-openai.html, accessed on 09.09.2024.
[440] *ChatGPT can now see, hear, and speak*, in *OpenAI*, 25 Sep. 2023, https://openai.com/index/chatgpt-can-now-see-hear-and-speak/, accessed on 09.09.2024.

«These features are part of an industrywide push toward so-called multimodal A.I. systems that can handle text, photos, videos and whatever else a user might decide to throw at them. The ultimate goal, according to some researchers, is to create an A.I. capable of processing information in all the ways a human can».[441] It is no coincidence that ChatGPT's advancements toward this direction have grown almost exponentially in a relatively narrow span of time, enabling users to engage in increasingly in-depth, natural, and consistent interactions, contributing to the consolidation of the chatbot's position among the most advanced GAI tools currently in existence.

In May 2024, GPT-4o (where the 'o' stands for 'omni') was released, a model capable of handling not only text but also images, audio and video, significantly faster and at half the cost of its predecessor, namely GPT-4 Turbo,[442] and with more advanced skills in natural language understanding and generation. Indeed, it achieved 88.7 in the Massive Multitask Language Understanding (MMLU) benchmark, specially formulated to assess the capabilities of LLMs and consisting of approximately 16 000 multiple-choice questions covering dozens of academic subjects such as math, medicine and philosophy, compared to the 86.4 achieved by GPT-4. In addition, GPT-4o «can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds, which is similar to human response time in a conversation».[443] Another significant improvement included the expansion of context windows, increased from earlier versions, permitting the model to retain more information in memory during a conversation or processing. GPT-4o also extended its capabilities beyond purely linguistic tests, successfully tackling

[441] K. Roose, *The New ChatGPT Can 'See' and 'Talk.' Here's What It's Like*, in «The New York Times», 27 Sep. 2023, https://www.nytimes.com/2023/09/27/technology/new-chatgpt-can-see-hear.html, accessed on 09.09.2024.

[442] *Hello GPT-4o*, in *OpenAI*, 13 May 2024, https://openai.com/index/hello-gpt-4o/, accessed on 04.09.2024.

[443] *Ibidem.*

benchmarks integrating multiple modalities such as image recognition and object classification, as well as showing improved learning efficiency with few-shots in tasks requiring advanced reasoning skills and multimodal data processing.[444] Furthermore, in July 2024, a smaller and more economical model, GPT-4o mini, was also launched. At present, this latter model is the default one for users who are not logged in and thus operate as guests, and for those who have not subscribed to the fee-based version and have depleted the (self-loading) limit of GPT-4o.

A further progress in terms of the capabilities of GAI systems was reportedly developed by OpenAI in September 2024, namely o1, a series of AI models «designed to spend more time thinking before they respond. They can reason through complex tasks and solve harder problems than previous models in science, coding, and math».[445] Central to this evolution is the implementation of large-scale reinforcement learning algorithms teaching the LLM to inherently employ Chain of Thought (CoT) to decompose problems. During fine-tuning, the model continuously refines these chains of thought. In terms of benchmarks, the o1-preview has shown remarkable performance in qualifying exams such as the American Invitational Mathematics Examination (AIME), and expert-level tests and quizzes in physics, chemistry, and biology (GPQA-diamond benchmark). OpenAI also developed and released o1-mini, a smaller and consequently computationally cheaper version optimised for generating and debugging code. These new models are subject to specific usage restrictions related to access and types of tasks

[444] SAKIB SHAHRIAR, ET AL. *Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency*, in *ArXiv*, 19 Jun. 2024, https://doi.org/10.48550/arXiv.2407.09519, accessed on 09.09.2024.

[445] *Introducing OpenAI o1-preview*, in *OpenAI*, 12 Sep. 2024, https://openai.com/index/introducing-openai-o1-preview/, accessed on 24.09.2024.

being supported according to the usage plan subscribed to with the company.[446] However, even these models, ostensibly humanised through statements during their reasoning such as «Thinking» or «Considering perspectives», are not exempt from the inherent limitations of GAI. Specifically, Apollo Research, an AI safety organisation aiming to mitigate potentially dangerous capabilities in advanced AI systems, recently noted how they are capable of deliberately lying, reiterating the phenomenon of fake alignment discussed in chapter I. The reason for this is the so-called 'reward hacking', emerging from the process of reinforcement learning, which can incentivise the model to produce answers that are not entirely correct but more rewarding for the user. In addition, o1 was valuated as a 'medium' risk for the ability to provide information to create biological or chemical weapons, although it does not allow non-experts to make threats on their own.[447] The deployment and development of ChatGPT marked a significant breakthrough in the field of AI, with far-reaching implications for multiple domains, including higher education, natural language capabilities augmented by techniques such as reinforcement learning from human feedback (RLHF), and the multimodal architecture implemented in the GPT-4 and GPT-4o models provide substantial benefits, while at the same time entailing significant challenges, both ethically and technologically. Notwithstanding advances in NLU, code generation, and interaction capabilities, LLMs persist in displaying numerous constraints, either inherent to the technology or resulting from the broader AI ecosystem: hallucinations, bias and lack of fairness, fake alignment, lack of transparency and interpretability, computational costs, resource intensive exploitation,

[446] *Learning to Reason with LLMs*, in *OpenAI*, 12 Sep. 2024, https://openai.com/index/learning-to-reason-with-llms/, accessed on 24.09.2024.

[447] KYLIE ROBISON, *OpenAI's new model is better at reasoning and, occasionally, deceiving*, in «The Verge», 17 Sep. 2024, https://www.theverge.com/2024/9/17/24243884/openai-o1-model-research-safety-alignment, accessed on 24.09.2024.

etc. Another limitation involves the escalating market pressure on companies developing GAI models, whose innovations seem to be motivated not just by scientific research requirements but also by commercial considerations and competitive pressing needs. For instance, in addition to the transition of OpenAI from a non-profit company, the introduction of a fee-based plan reflects a rising tendency toward monetisation of GAI. This shift towards market targets might influence the type of research that is financed and promoted, with the risk of overshadowing alignment and security concerns in favour of more profitable solutions in the short term. Future choices made by the major stakeholders will not only determine technological progress, but also how society will manage the potential risks associated with this powerful technology. A balance between innovation, ethical responsibility and sustainability, in the broadest sense of the term, will therefore continue to be crucial.

ChatGPT therefore constitutes a revolutionary approach in terms of the accessibility of GAI and LLMs, and is currently being used in a wide variety of practical scenarios due to its generalisability.[448] Its employment in this pilot study was dictated precisely by the popularity of the tool, which is regularly utilised by a large number of university students.[449] Therefore, it appeared necessary to employ a tool with which they were familiar, while at the same time allowing them to become thoroughly acquainted with its potential, limitations and implications, then informing their subsequent interactions.

[448] JINGFENG YANG, ET AL., *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*, in «ACM Transactions on Knowledge Discovery from Data», vol. 18, n. 160, pp. 1-32, https://doi.org/10.1145/3649506, accessed on 28.09.2024.

[449] ERIC ROSENBAUM, *AI is getting very popular among students and teachers, very quickly*, in «CNBC», 11 Jun. 2024, https://www.cnbc.com/2024/06/11/ai-is-getting-very-popular-among-students-and-teachers-very-quickly.html, accessed on 28.09.2024.

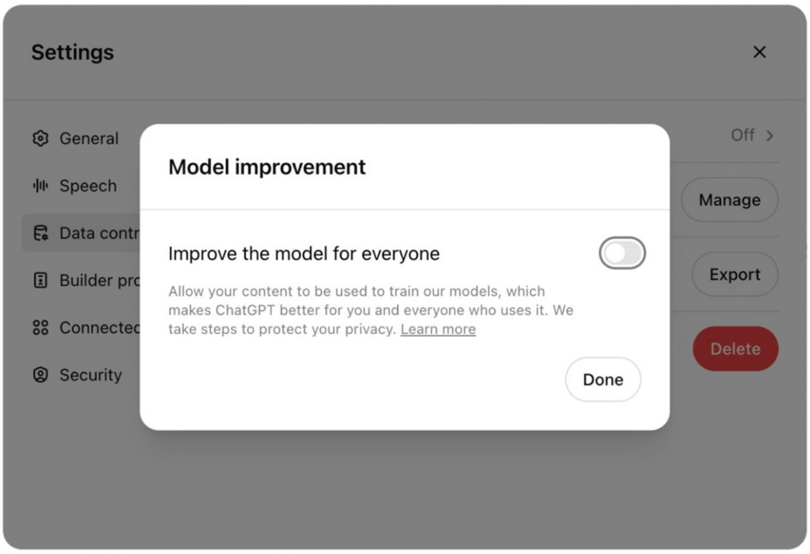
SIXTH CHAPTER

DESIGN AND METHODOLOGIES OF THE EMPIRICAL STUDY

VI.1. Methodology

A pilot study for the purposes of this thesis, i.e. assessing the effect of an AIL workshop with a specific focus on GAI, and LLMs-based systems such as ChatGPT, was conducted with undergraduate students from different humanities faculties (Philosophy, International, and Economic Studies; Language, Civilisation and the Science of Language; Conservation of Cultural Heritage and Performing Arts Management; Economics and Management of Arts and Cultural Activities) at Ca' Foscari University of Venice. It culminated in a three-and-a-half-hour long workshop held from 9:30 a.m. to 1:00 p.m. on 6th September 2024 and taught by the present author with the invaluable support of Professor Teresa Scantamburlo of the Department of Environmental Sciences, Informatics and Statistics. The workshop was hosted within one of the venues provided by the university for events and conferences, Aula Tesa 1 of the CFZ library, Ca' Foscari Zattere. The room was chosen as it offered a computer connected to a projector.

Prior to starting the research, formal approval was requested from the University Ethics Committee regarding the processing of students' personal data: on 10 June 2024, the Committee issued a favourable opinion. In fact, privacy and data protection were given the highest priority throughout all phases of the pilot study, considering the implications of data transmission to OpenAI, students' informed consent, and raising awareness about these concerns during the workshop. As for the questionnaires, they were administered through Google Forms. Participants were expected to access them through their UniVe account for security reasons, however their e-mail addresses were not collected and therefore the answers were anonymous. Measures taken to safeguard their privacy also comprised the restriction of access to results only to members of the research team; in addition, non-traceability of ChatGPT interactions to student identities was ensured.



Model improvement

Improve the model for everyone

Allow your content to be used to train our models, which makes ChatGPT better for you and everyone who uses it. We take steps to protect your privacy. Learn more

FIGURE 17, *Screenshot showing how to disable the transmission of personal data and data collected during ChatGPT interactions to OpenAI.*

Before starting the workshop, participants were explained how to disable the utilisation of their data for the improvement (fine-tuning) of ChatGPT and it was ensured that everyone deactivated this option on their device (Figure 17). Concurrently, in order to identify the key focus points of the study, opinions were sought from professors of the aforementioned faculties, from experts, and from students, so as not to overlook instances of the use of GAI in academic environments and critical issues that would later be addressed within the literacy workshop. This confrontation phase was crucial, as it allowed to provide more scientific rigour to the pilot study design and proposed methodologies.

The workshop was conducted as an informal and educational initiative on GAI and prompt engineering (Figure 18). A questionnaire together with a ChatGPT exercise was given to the participants both before and after the AIL informative part, in order to assess how they were interacting, what their opinions and knowledge were, and to measure a possible shift between the before and after. As for the literacy part, it was divided into two modules of 45 minutes each with a short break in between. The first module focused on explaining the definition of 'intelligence' according to a teleological and artificial intelligence perspective, a brief history of AI and GAI (section I.1) and LLMs, a brief overview of the architecture of LLMs and ChatGPT, the intrinsic limitations of these models (section I.2.1 and I.2.2), the ethical, legal and social implications of GAI (section I.4), as well as successful use cases for AI for good. In the second part, instead, the concept of prompt engineering was addressed, and the CLEAR framework for effective input formulation was introduced, along with several prompt engineering techniques and prompt patterns (chapter III) consisting in live hands-on practice with ChatGPT in order to maximise participants' engagement.[450] The workshop content and activities were

[450] S.-C. KONG, W. MAN-YIN CHEUNG, G. ZHANG, *Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds*, cit.
E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

designed taking into consideration a literature review of already developed studies in this area.[451]

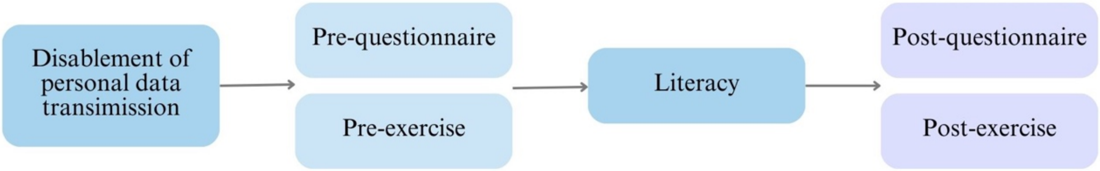
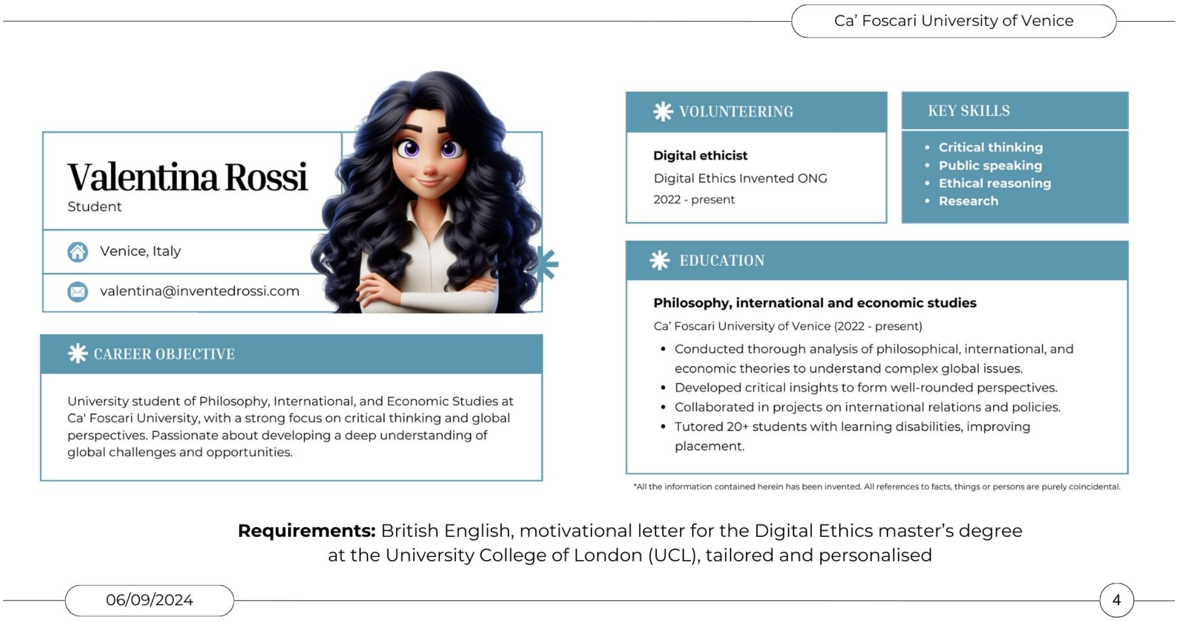


FIGURE 18, *Workshop phases.*

In support of the learning, informative slides were designed and shared with the students after the workshop. At the conclusion, a short session was held to answer questions and collect participants' general feedback. The questionnaire questions and the material relating to the ChatGPT exercise can be found in the appendix to this thesis. More specifically, the prompt engineering exercise required the students to have the chatbot generate a motivational letter for a potential Master's degree application. The

[451] N. KNOTH, ET AL., *AI literacy and its implications for prompt engineering strategies*, cit.
S.-C. KONG, W. MAN-YIN CHEUNG, G. ZHANG, *Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds*, cit.
L. S. LO, *The CLEAR path: A framework for enhancing information literacy through prompt engineering*, cit.
MARITA SKJUVE, ASBJØRN FØLSTAD, PETTER BAE BRANDTZAEG, *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*, in «CUI '23: ACM conference on Conversational User Interfaces», 2023, https://doi.org/10.1145/3571884.3597144, accessed on 10.04.2024.
MIRIAM SULLIVAN, ET AL., *Improving students' generative AI literacy: A single workshop can improve confidence and understanding*, in «Journal of Applied Learning & Teaching», vol. 7, n. 2, 2024, http://journals.sfu.ca/jalt/index.php/jalt/index, accessed on 10.09.2024.
E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.
YOSHIJA WALTER, *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*, in «International Journal of Educational Technology in Higher Education», vol. 21, n. 15, 2024, https://doi.org/10.1186/s41239-024-00448-3, accessed on 21.07.2024.
DAVID JAMES WOO, ET AL., *Effects of a Prompt Engineering Intervention on Undergraduate Students' AI Self-Efficacy, AI Knowledge, and Prompt Engineering Ability: A Mixed Methods Study*, in *ArXiv*, 30 Jul. 2024, https://doi.org/10.48550/arXiv.2408.07302, accessed on 20.08.2024.

letter had to meet a set of instructions to be contained as output of the interaction, i.e. the use of British English, the absence of hallucinations, and the structure of a letter in terms of format (Figure 19). Students could interact with up to 5 prompts within the same conversation with the tool. This same exercise was repeated after the explanation of the prompt engineering techniques and prompt patterns. Both in the pre and post-questionnaire there was a section where it was possible to enter the shared links of the chats (after anonymising them, following the instructions given in the assignment).



Ca' Foscari University of Venice

Valentina Rossi

Student

Venice, Italy

valentina@inventedrossi.com

CAREER OBJECTIVE

University student of Philosophy, International, and Economic Studies at Ca' Foscari University, with a strong focus on critical thinking and global perspectives. Passionate about developing a deep understanding of global challenges and opportunities.

VOLUNTEERING

Digital ethicist

Digital Ethics Invented ONG

2022 - present

KEY SKILLS

- **Critical thinking**
- **Public speaking**
- **Ethical reasoning**
- **Research**

EDUCATION

Philosophy, international and economic studies

Ca' Foscari University of Venice (2022 - present)

- Conducted thorough analysis of philosophical, international, and economic theories to understand complex global issues.
- Developed critical insights to form well-rounded perspectives.
- Collaborated in projects on international relations and policies.
- Tutored 20+ students with learning disabilities, improving placement.

*All the information contained herein has been invented. All references to facts, things or persons are purely coincidental.

Requirements: British English, motivational letter for the Digital Ethics master's degree at the University College of London (UCL), tailored and personalised

06/09/2024

4

FIGURE 19, *Material for the prompt engineering exercise.*

VI.2. Selection of participants

Participants were selected on a voluntary basis, through an open call disseminated via the official email channel of the Philosophy, International, and Economic Studies faculty and via several social networks, such as WhatsApp Messenger and Instagram. The only requirement for participation was the possession of an active OpenAI account, so as to be able to access and interact with ChatGPT during the workshop and questionnaire exercises. No additional criteria for inclusion or exclusion were imposed, since the primary objective was to assess the impact of literacy in GAI and prompt engineering techniques on a heterogeneous sample of humanities students. The internal diversity of the sample in terms of background, interests and future perspectives contributed to provide a nuanced representation of ChatGPT experiences and use cases within this academic context. The sample consisted of 9 undergraduate students, and its demographic characteristics included a representation of 22.22% male and 77.78% female students. The age distribution resulted in 22.22% of the participants being between 18 and 20 years old, 66.67% between 20 and 22 years old, and 11.11% between 22 and 24 years old. 100% of the participants were of European nationality.

The relatively low number of participants is justified by this study being a pilot phase: in this instance, the restricted sample allowed for significant preliminary results, while reducing the cost and time required to perform the research, and providing an initial glimpse of the impact this initiative is likely to have on a specific group of university students. Furthermore, this initial phase allowed the identification and resolution of practical problems, the evaluation of participants' responsiveness and the validity of the data collection tools used. The pilot results will provide crucial data to possibly modify and improve the study design, ensuring the proposed methodologies are effective and applicable on a larger scale. A possible self-selection bias must be acknowledged: the participants, having voluntarily chosen to participate in this pilot study, might have been more involved in artificial intelligence or ChatGPT than average students in general, and this factor could influence the findings, since they might therefore also be more likely to use ChatGPT positively or to evaluate the prompt engineering techniques suggested during the workshop more favourably. As a result, the self-selection bias could limit the generalisability of the study results, rendering it more difficult to extrapolate the findings to a wider population of humanities students or different academic contexts. In order to mitigate the impact of this bias, in future stages of the research it might be useful to consider the adoption of a recruitment methodology employing a more random or stratified sampling, possibly including students who do not show a particular interest in artificial intelligence. The decision to engage humanities students addresses specific needs: these students, as opposed to those from scientific or technological disciplines, are generally less familiar with artificial intelligence systems such as LLMs,[452] especially from a technical and operational point of view, thereby making them an ideal group to explore how an introduction to these technologies might influence their understanding and use of the tools. Furthermore, the development of a critical and nuanced thinking attitude generally resulting from the study of humanities subjects also enables these students to be particularly attentive to ethical, cultural and social aspects, which have formed and should form an essential part of any artificial intelligence literacy course.[453] «If the technology is going to be directed in a more socially responsible way, it is time to dedicate time and attention to AI ethics education».[454] This approach responds to the increasingly acknowledged urgency to devote attention to educating on AI ethics in order to steer technological development in a socially responsible manner.

[452] SIU-CHEUNG KONG, WILLIAM MAN-YIN CHEUNG, GUO ZHANG, *Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds*, in «Computers and Education: Artificial Intelligence», vol. 2, 2021, https://doi.org/10.1016/j.caeai.2021.100026, accessed on 05.06.2024.

[453] ANDREA ALER TUBELLA, MARÇAL MORA-CANTALLOPS, JUAN CARLOS NIEVES, *How to teach responsible AI in Higher Education: challenges and opportunities*, cit.

[454] JASON BORENSTEIN, AYANNA HOWARD, *Emerging challenges in AI and the need for AI ethics education*, in «AI and Ethics», vol.1, 2021, pp. 61-65, https://doi.org/10.1007/s43681-020-00002-7, accessed on 15..05.2024.

VI.3. Questionnaire design

With regard to the questionnaire, some of the questions were adapted from previous studies in order to provide scientific rigour and a solid grounding and others, specifically those in the macro-sections *Perceived interaction improvement*, and *Ethical considerations*, were purposely designed for the purposes of this pilot study: the first to directly monitor participants' feedback, the latter to measure a pre and post-workshop possible shift concerning ethical and pragmatic awareness of GAI and prompt engineering. The decision to adopt a Likert scale for the questionnaire design is based on a series of methodological considerations that have made this technique a popular instrument for the measurement of opinions, attitudes and behaviours in the fields of social and educational research. Indeed, the feasibility of quantifying perceptions in such a way that they can be easily analysed using statistical methods is a major advantage that led to its endorsement.[455] The typical Likert scale presents a set of statements to which respondents express their degree of agreement or disagreement on a numerical scale, generally ranging from five to seven points. However, the design of this questionnaire employed a five-point scale as it allowed for greater simplicity for respondents, minimising confusion and ambiguity in the answer, thereby avoiding the effort that can arise with longer scales.[456]

The first section, *Previous experience with AI, GAI and ChatGPT*, was created expressly in order to verify the students' level of previous experience and knowledge with

[455] TAKASHI YAMASHITA, ROBERTO J. MILLAR, *Likert Scale*, in DANAN GU, MATTHEW E. DUPRE (edited by), *Encyclopedia of Gerontology and Population Aging*, Cham, Springer, 2021, https://doi.org/10.1007/978-3-030-22009-9_559, accessed on 20.03.2024.
[456] *Designing Likert scales*, in *TASO*, https://taso.org.uk/evidence/evaluation-guidance-resources/survey-design-resources/evaluation-guidance-designing-likert-scales/, accessed on 20.03.2024.

regard to these technologies: the questions focused on investigating familiarity with concepts such as artificial intelligence, generative artificial intelligence, and prompt engineering, the answers to which consisted of a scale from '*No knowledge*' (1) to '*Very good knowledge*' (5), as well as prior experience with ChatGPT (both the basic version and the premium, paid version) generally and within the academic context, the answers to which consisted of a scale from '*Never*' (1) to '*I use it almost every day*' (5). This question section was only included in the pre-questionnaire.

The second macro-section, i.e. *Perceived level of threat*, was adapted from a previous study conducted and related to an AI literacy workshop in an Italian secondary school, which itself tailored the questions from an earlier one concerning the exploration of the effects of robots' anthropomorphism and of how their capabilities affect the perception of threat.[457] The results suggest how robots with a high level of anthropomorphism (androids), particularly those seemingly superior to humans in terms of skills and abilities, are perceived as a threat not only to human security and resources (realistic threat), but also to the identity and uniqueness intrinsically constituting human nature (identity threat). This study highlights the crucial importance of considering such an effect in robot design. However, as discussed in chapter II, the perception of anthropomorphism, and in particular the ELIZA effect, are also particularly poignant issues for GAI systems, therefore the integration of this type of analysis within the questionnaire seemed pertinent. This set of questions was part of both the pre and post questionnaire and the answers consisted of a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5).

[457] KUMAR YOGEESWARAN, ET AL., *The Interactive Effects of Robot Anthropomorphism and Robot Ability on Perceived Threat and Support for Robotics Research*, in «Journal of Human-Robot Interaction» vol. 5, n. 29, http://dx.doi.org/10.5898/JHRI.5.2.Yogeeswaran, accessed on 10.04.2024.

The following section, *Emotions resulting from interaction*, was also adapted from the aforementioned study,[458] which itself adapted it from a major framework used to quantify discrete emotions, namely the Discrete Emotions Questionnaire (DEQ),[459] focusing on eight specific emotions: anger, disgust, fear, anxiety, sadness, happiness, serenity and desire. Each emotion is supported by solid theories on the underlying mechanisms: for instance, anger is described as a highly-activated, and negative emotion, related to an attachment motivation, whereas disgust is a highly-activated, and negative emotion, but with an avoidance motivation. Students were asked to measure the degree of their emotions experienced during interactions with ChatGPT on a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5). This section was only included in the post-questionnaire.

The fourth section, i.e. *Interaction quality evaluation*, was adapted from the high school workshop study, and was intended to include measures of the perceived functionality of ChatGPT and the effort required to obtain the desired behaviour from the tool. Also this set of questions was part of both the pre and post questionnaire and the answers consisted of a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5).

The fifth section, *ChatGPT as a tool*, saw some questions adapted from two previous studies: specifically, the sub-section *Pragmatic dimension*, the questions '*ChatGPT is exceeding expectations, impressive, or superior compared to existing solutions*', '*ChatGPT can support creative activities (such as essay writing,*

[458] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

[459] CINDY HARMON-JONES, BROCK BASTIAN, EDDIE HARMON-JONES, *The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions*, in «PloS one», vol. 11, n. 8, 2016, https://doi.org/10.1371%2Fjournal.pone.0159915, accessed on 07.04.2024.

brainstorming or dialectical exchanges)', and '*Interactions with ChatGPT are entertaining'* from the sub-section *Hedonic dimension*, and the question '*ChatGPT is humanlike or intelligent*' from the sub-section *Human likeness* were adapted from a questionnaire developed to analyse the user experience with ChatGPT using the pragmatic-hedonic framework, distinguishing between pragmatic attributes, which relate to the usefulness and productivity of the tool, and hedonic attributes, which include entertainment and creative interactions.[460] The questions in the *Social presence* section, on the other hand, were adapted from the pilot study on the AIL workshop in a high school mentioned above.[461] The questions that have not been mentioned so far were purposely crafted for this study. This macro-section was repeated in both the pre and post-questionnaire, and the answers consisted of a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5).

The following section, i.e. *Perceived interaction improvement*, was designed in order to measure students' perceptions of personal improvement after the workshop was held: it was only included in the post-questionnaire, and these answers also consisted of a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5).

Parallel to this, the last macro-section, *Ethical considerations*, was also created specifically within the context of this empirical study. The questions were intended to measure the participants' awareness and opinion regarding both the ethical and legal threats of GAI and AI more generally, as well as a future usefulness dimension of prompt engineering and AI literacy. This macro-section was repeated in both the pre and post-

[460] M. SKJUVE, A. FØLSTAD, P. BAE BRANDTZAEG, *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*, cit.

[461] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

questionnaire, and the answers consisted of a Likert scale ranging from *Strongly disagree* (1) to *Strongly agree* (5).

VI.4. Evaluation criteria

Regarding the evaluation of the interactions, both the prompts provided by the students as well as the resulting outputs were evaluated in order to understand how the prompt engineering techniques introduced during the workshop influenced the students' ability to formulate precise and optimised prompts for LLMs and obtain relevant results.

In the first phase, a mixed analysis of the prompts was carried out, integrating the interpretation and categorisation of the inputs employed by the students (both before and after the presentation of prompt engineering techniques) through a qualitative approach, evaluating them on a scale of 1 to 5. Specifically, the aspects related to the CLEAR framework that was proposed as a possible guidance for the formulation and iteration of effective prompts were assessed:[462] clarity, logicality, explicitness, adaptability and reflexivity. The decision to adopt this framework both for AIL and for the evaluation of formulated requests was motivated by the presence, notably, of the 'adaptive' and 'reflective' components: they allow users to maintain a central role in supervising, correcting and improving their interactions with LLMs and GAI while maintaining control and validation over the results they generate and the requests they submit. This enables the evaluation of prompts not only as stand-alone, but also as an active part of collaboration between humans and artificial intelligence systems. In a human-in-the-loop model of interaction, as detailed in chapter II, both these principles are vital to ensure that technology is not a delegate of one's decision-making faculty, or an autonomous system,

[462] L. S. Lo, *The CLEAR path: A framework for enhancing information literacy through prompt engineering*, cit.

but rather functions as an empowering extension of human expertise. Concerning prompts, quantitative metrics were also considered, such as the number of interactions with the chatbot within the same conversation (i.e. the number of input/output pairs), the presence or absence of attached files within the prompt (i.e. the fictional curriculum vitae included within the delivery), and the total number of input words.

In the second phase, a mixed analysis was performed as well, in this case concerning the outputs, by measuring on a scale of 1 to 5 the adherence to the requirements included in the delivery of the prompt engineering exercise: respectively, the absence of hallucinations, adherence to the language requirement (i.e., British English), adherence to the format (in terms of the structure of a letter), and originality. Again, the quantitative metric of the number of words in the output was taken into account. The two phases were repeated with the same methodologies for both pre-questionnaire and post-questionnaire interactions: the data were then cross-referenced to understand the variation in interactions before and after the workshop and the explanation of how GAI and prompt engineering and prompt patterns operate.

In sum, the questionnaire, which was structured into several macro-sections, enabled data to be collected on several dimensions, such as previous experience, perceived threat, emotions raised by interactions and the perceived quality of the interactions themselves. A five-point Likert scale was employed to ease the collection of measurable responses, allowing for a quantitative analysis of opinions and attitudes. At the same time, the qualitative and quantitative evaluation of the prompts and outputs provided a thorough insight into the impact of the prompt engineering techniques introduced during the workshop.

SEVENTH CHAPTER

ANALYSIS OF THE RESULTS OF THE EMPIRICAL STUDY

VII.1. Data analysis

The data collected in the pilot study were analysed statistically in order to derive knowledge, and to assess the effectiveness of the workshop in imparting knowledge on GAI, LLMs and prompt engineering techniques. Data analysis of the questionnaires was carried out using descriptive statistics and inferential statistical tests: in particular, each response on a Likert scale was coded numerically, varying from 1 (*Strongly disagree*) to 5 (*Strongly agree*). Each question's data was abstracted both by calculating the mean (summing the numerical values and dividing them by the number of answers) and the standard deviation, which serves as a measure of variability around the mean to capture central tendencies and dispersion between answers, respectively. The evaluation of prompts and outputs, on the other hand, was based on qualitative and quantitative metrics. Qualitatively, prompts were analysed based on the CLEAR framework focusing on clarity, logicality, explicitness, adaptability and reflexivity. The outputs were assessed based on the predetermined criteria of absence of hallucinations, adherence to British English, correct letter format and originality. Ratings were given on a scale of 1 to 5 for each criterion, with 5 indicating excellent alignment with the desired characteristics.

Additionally, the total number of prompt-output pairs in each session was counted to measure user engagement, as well as the number of input and output words to assess conciseness and expansiveness of communication. Subsequently, a comparative analysis of the mean scores of responses and prompt and output evaluations to the pre and post questionnaires was performed in order to reveal changes in knowledge, attitudes and perceptions regarding GAI and prompt engineering.

Besides, to estimate the statistical significance of changes, the Wilcoxon matched-pairs test was employed, after verifying the normality of the distributions with the Shapiro-Wilk test, for which most categories showed very low p-values (< 0.05), indicating how the distribution of the data actually did not follow a normal pattern. Therefore, given the nature and features of the data collected, it was necessary to use the Wilcoxon test for several reasons: namely, this non-parametric test is particularly suitable for paired and non-normal data such as those collected, comprising responses on Likert scales (using the Student t-test, normality of the distribution would have been assumed instead). This test appears to be the most appropriate as it compares rank averages rather than absolute values, reducing the distorting effect of outliers and ensuring the reliability of results even with small samples. If the p-value is less than or equal to 0.05 (often referred to as $\alpha = 0.05$), results are considered to be statistically significant. Analyses were performed using Python, and in particular, using the `shapiro` and `wilcoxon` from the library `scipy.stats`. The `mode='exact'` parameter was used due to the small sample size analysed: this approach provided a more accurate estimate of statistical significance, calculating exact probabilities based on all possible permutations of the pre and post-questionnaire responses.

Previous experience with AI, GAI and ChatGPT (pre-questionnaire)

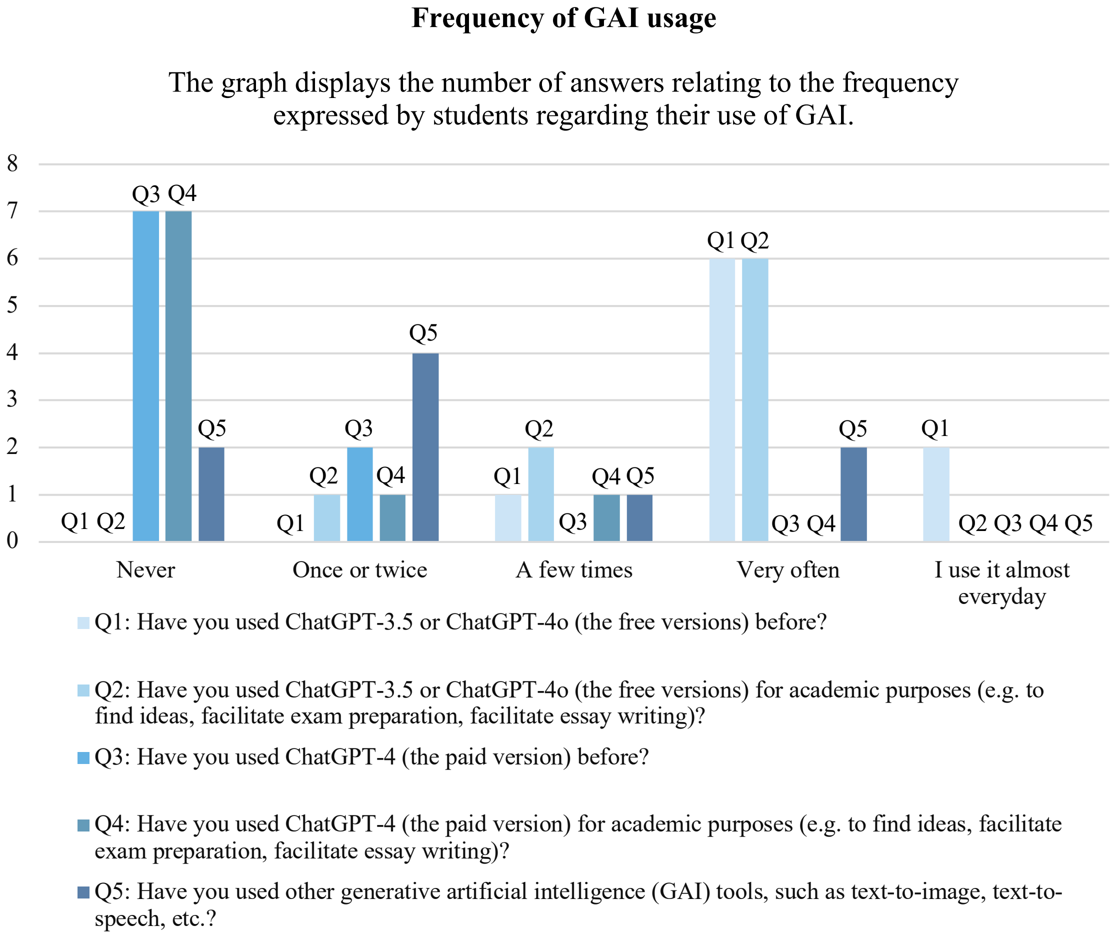


FIGURE 20, *Frequency of GAI usage.*

The first section of the questionnaire aimed to understand what level of knowledge and expertise the students had towards AI and GAI: this is important information to understand how to tailor educational material for possible further studies. The majority of the students reported having '*Limited knowledge*' of the concept of artificial intelligence (55.56%) and generative artificial intelligence (66.67%) and '*No knowledge*' of prompt engineering (55.56%). Likewise, the respondents reported using the free version of ChatGPT, i.e. ChatGPT-3.5 until May 2024 and ChatGPT-4o, '*Very often*' (66.67%) both in general and for academic purposes (e.g. to find ideas, facilitate exam preparation, facilitate essay writing). As for the fee-based premium version, i.e. ChatGPT-4, it was '*Never*' used by the majority of students (77.78%) both in general and in a university context. A considerable number of students (77.78%), on the other hand, used at least once other GAI systems, such as text-to-image or text-to-speech: the ones most frequently mentioned were DALL·E within Microsoft Bing Image Creator[463] and Midjourney,[464] both two of which are intended to generating images from textual prompts.

Perceived level of threat (pre-questionnaire, post-questionnaire)

The perception of realistic and identity threat in relation to AI decreased after the workshop, analogous to the previous study:[465] mPre = 3.08, sd = 1.12; mPost = 2.56, sd = 1.00. The Wilcoxon test performed for each question did not show any statistically significant changes, except for the question '*The realism of artificial intelligence is disturbing because it makes it almost indistinguishable from human beings*' (mPre = 3.33, sd = 1.25; mPost = 2.44, sd = 0.83) for which the p-value of 0.054 is close to the threshold, suggesting a change in perception: this concern was perceived as less impactful after the workshop. The results of the other questions were: '*Artificial intelligence applications are beginning to blur the boundaries between human and machine*' (mPre = 3.33, sd = 0.94; mPost = 2.56, sd = 0.68); '*The increased use of artificial intelligence in our lives is causing humans to lose their jobs*' (mPre = 2.89, sd = 1.29; mPost = 2.89, sd = 0.99); '*Artificial intelligence implementations can effectively replace workers from their jobs*' (mPre = 3.22, sd = 0.92; mPost = 2.67, sd = 1.05); '*In the long run, artificial intelligence*

[463] https://www.bing.com/images/create/

[464] https://www.midjourney.com

[465] E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

poses a direct threat to human welfare and safety' (mPre = 3.33, sd = 0.94; mPost = 2.67, sd = 1.05); '*Recent advances in Artificial Intelligence are challenging the very essence of what it means to be human*' (mPre = 2.67, sd = 1.15; mPost = 2.11, sd = 0.99); '*Technological advances in artificial intelligence are threatening the uniqueness of humans*' (mPre = 2.78, sd = 1.03; mPost = 2.56, sd = 1.17).

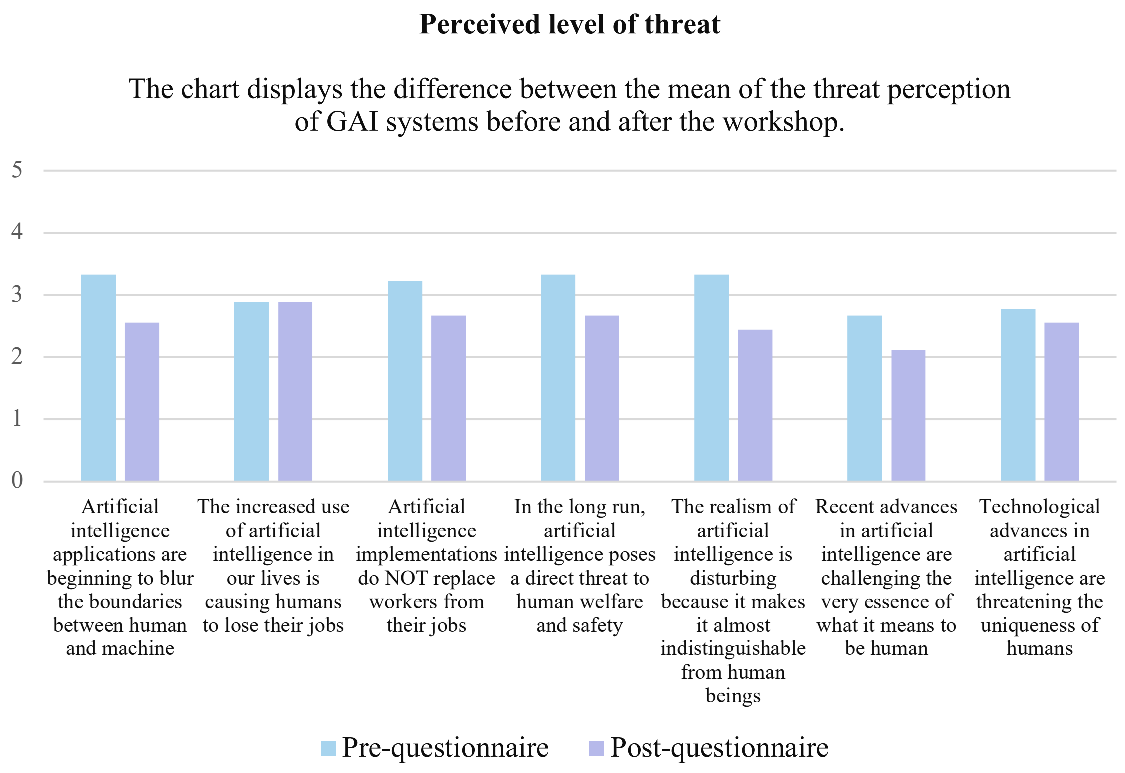


FIGURE 21, *Perceived level of threat.*

The workshop was generally effective on the perception of AI and GAI-related threats: after the session, participants showed a decrease in their concern regarding these issues. Nevertheless, it was only for the specific issue related to how AI's realism might challenge human beings, that approaching statistical significance was observed: this suggests that, whilst the AIL initiative successfully influenced general attitudes towards

AI, some deep-rooted preoccupations are still difficult to modify: the workshop discussions may have offered new perspectives that contributed to mitigating some fears, but it is clear that dialogue needs to continue in order to address more grounded concerns.

Emotions resulting from interaction (post-questionnaire)

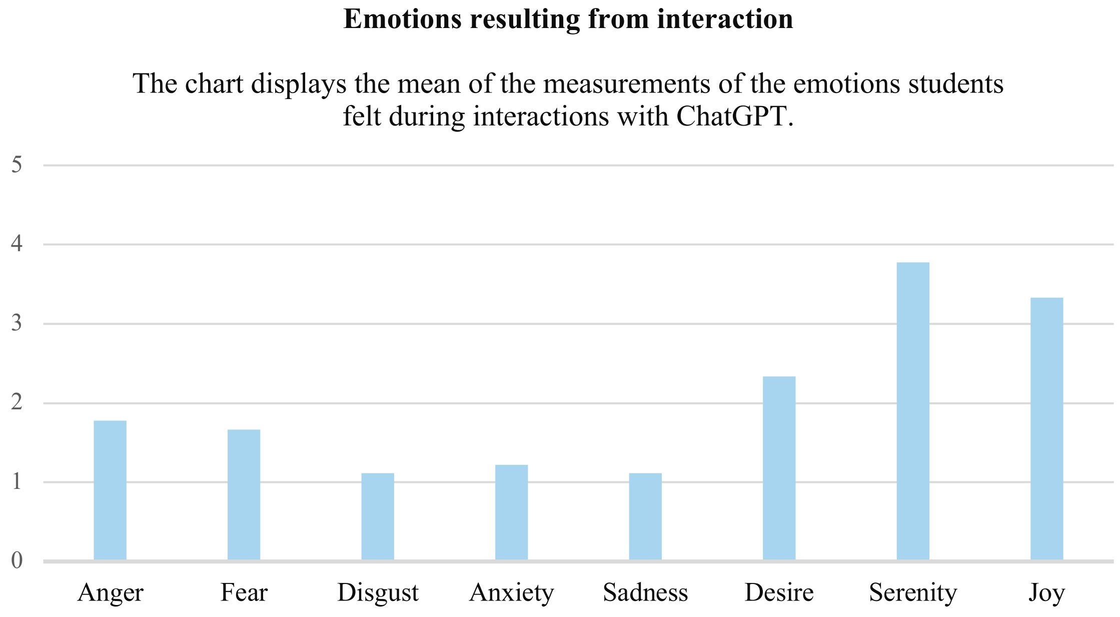


FIGURE 22*, Emotions resulting from interaction.*

The students reported significantly higher positive emotions than negative ones: the emotion with the highest mean was '*Serenity'* (m = 3.78, sd = 0.79). The results for the other emotions were: '*Anger*' (m = 1.78, sd = 0.79), '*Fear*' (m = 1.67, sd = 1.25), '*Disgust*' (m = 1.11, sd = 0.31), '*Anxiety*' (m = 1.22, sd = 0.41), '*Sadness*' (m = 1.11, sd = 0.31), '*Desire*' (m = 2.33, sd = 1.05), and '*Joy*' (m = 3.33, sd = 0.67). These data indicate that participants predominantly experienced positive emotions during their interactions with ChatGPT, suggesting that the dialogue experience is perceived as reassuring or positive rather than a source of anxiety or fear. The high occurrence of emotions such as joy also denotes an enthusiastic reception towards technology, reflecting a possible decrease in resistance towards the adoption of such tools in everyday life: this scenario is consistent with the findings of the frequency of GAI usage in the first section. A conversational evaluation carried out at the end of the workshop, during the question time, revealed how the clean interface of the tool, which is not perceived as confusing and appears to have been accessible and effective for the participants, plays a significant influence.

Interaction quality evaluation (pre-questionnaire, post-questionnaire)

The evaluation of the interactions with ChatGPT revealed a slight improvement in the post-workshop phase compared to the pre-questionnaire exercise: mPre = 3.17, sd = 1.06; mPost = 3.19, sd = 1.06. Specifically, with regard to ChatGPT's capabilities, the question '*I am satisfied with ChatGPT's comprehension and response capabilities*' (mPre = 3.33, sd = 0.47; mPost = 4.00, sd = 0.47), recorded a p-value of 0.014 according to the Wilcoxon test, thus indicating a statistically significant improvement in the satisfaction of the chatbot's comprehension and response capabilities following the introduction of the prompt engineering and prompt patterns techniques. The results of the other questions were: '*ChatGPT demonstrates to be intelligent*' (mPre = 3.44, sd = 0.83; mPost = 3.56, sd = 0.68), '*ChatGPT repeats the same mistakes over and over again, without adapting to my questions*' (mPre = 2.78, sd = 1.03; mPost = 2.33, sd = 0.67), '*ChatGPT demonstrates an understanding of complex concepts*' (mPre = 3.11, sd = 1.10; mPost = 3.00, sd = 0.82), '*ChatGPT demonstrates human-like reasoning and comprehension*' (mPre = 2.89, sd = 0.87; mPost = 2.56, sd = 1.07).

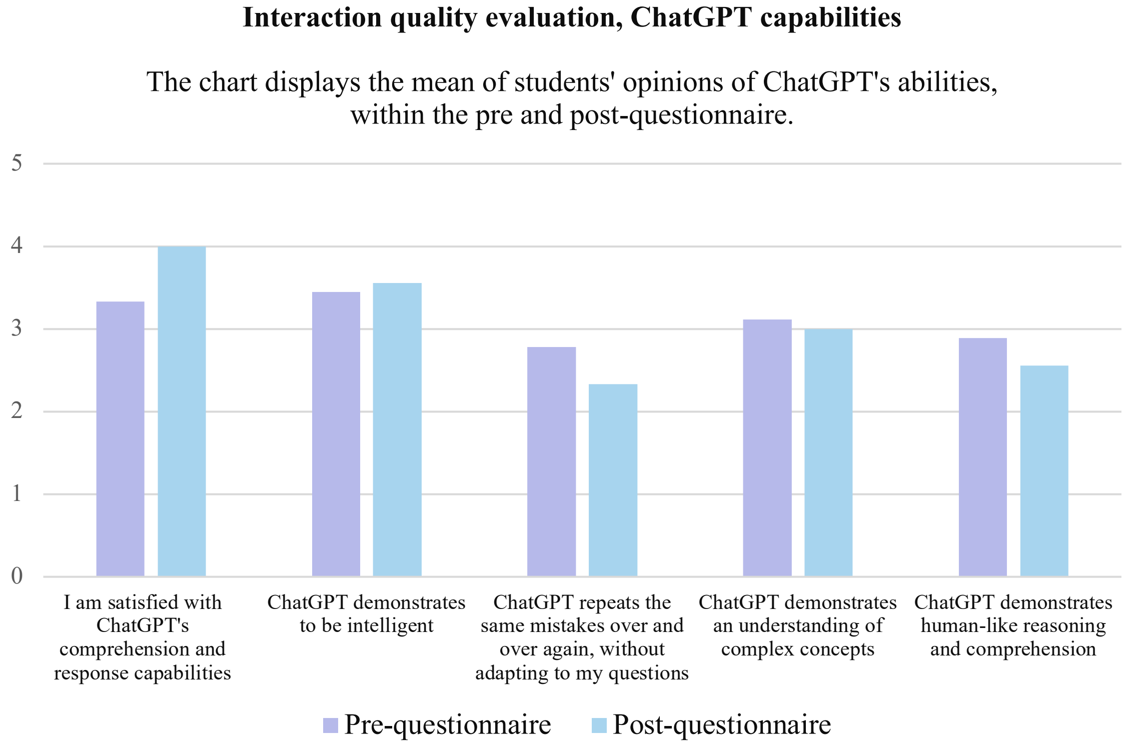


FIGURE 23*, Interaction quality evaluation, ChatGPT capabilities.*

With regard to the effort perceived to obtain the intended output and behaviour, improvements were also registered – although the Wilcoxon test did not identify any statistically relevant ones. The results of the questions were: '*I found it easy to communicate my intentions to ChatGPT*' (mPre = 3.44, sd = 1.07; mPost = 4.11, sd = 0.57), '*Obtaining desired responses from ChatGPT required an acceptable level of effort on my part*' (mPre = 3.11, sd = 1.20; mPost = 3.11, sd = 1.20), '*I had to repeat my questions or requests multiple times to get satisfactory responses from ChatGPT*' (mPre = 3.33, sd = 0.94; mPost = 3.00, sd = 0.82), '*Interacting with ChatGPT required more effort than I initially expected*' (mPre = 3.11, sd = 1.52; mPost = 3.00, sd = 1.33). The analysis revealed an improvement in student satisfaction with the model's comprehension and response capabilities, and while enhancements in other areas were not statistically significant, the results indicate a positive trend towards increased ease of communication and less effort required to obtain desired responses, pointing out the efficacy of the workshop and prompt engineering techniques.

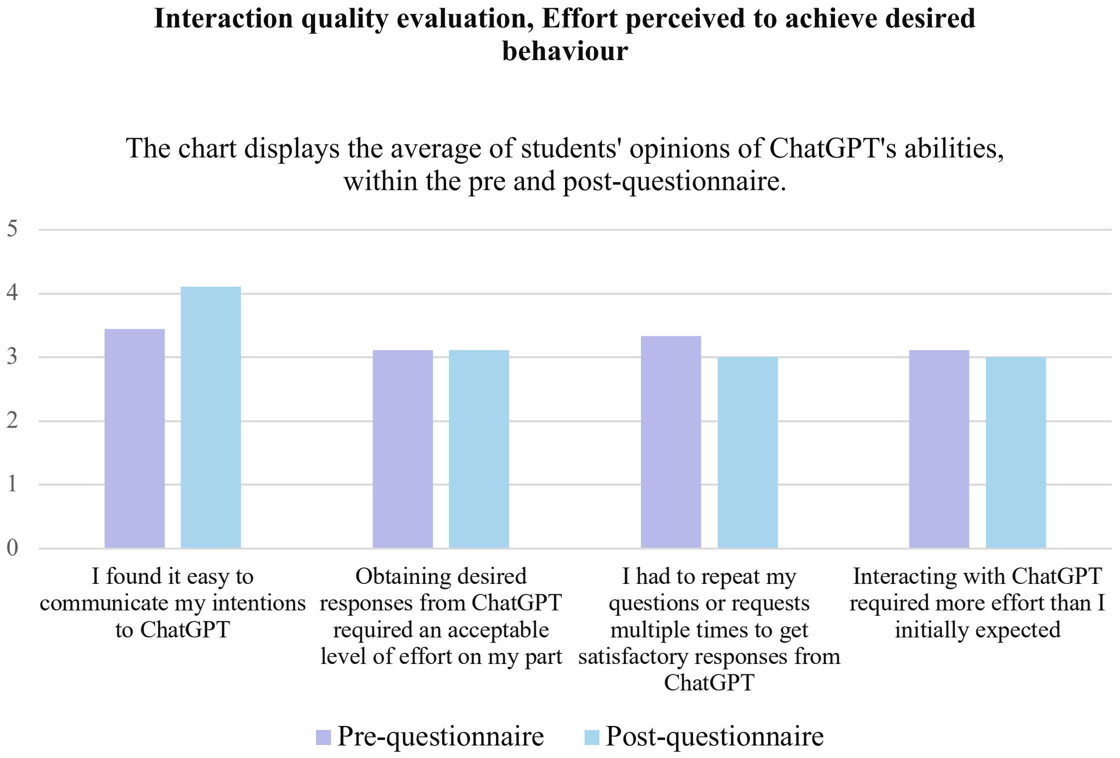


FIGURE 24, *Interaction quality evaluation, Effort perceived to achieve desired behaviour.*

ChatGPT as a tool (post-questionnaire)

Participants generally showed a positive opinion of ChatGPT, highlighting how it enhances work efficiency and produces useful and relevant outputs: m = 3.31, sd = 1.71. Specifically, the results of the questions regarding the *Pragmatic dimension*, i.e. the evaluation of functional, utilitarian and practical aspects of the tool, focusing on how effective it is at accomplishing what it promises, were: '*ChatGPT enables efficiency and enhances quality of work*' (m = 4.44, sd = 0.50), '*ChatGPT provides relevant and useful* *output*' (m = 4.33, sd = 0.47), '*ChatGPT does not want to answer the request due to policies or because it only has limited information in its database*' (m = 1.89, sd = 0.74), '*ChatGPT presents misinformation or biased views*' (m = 2.56, sd = 1.07).

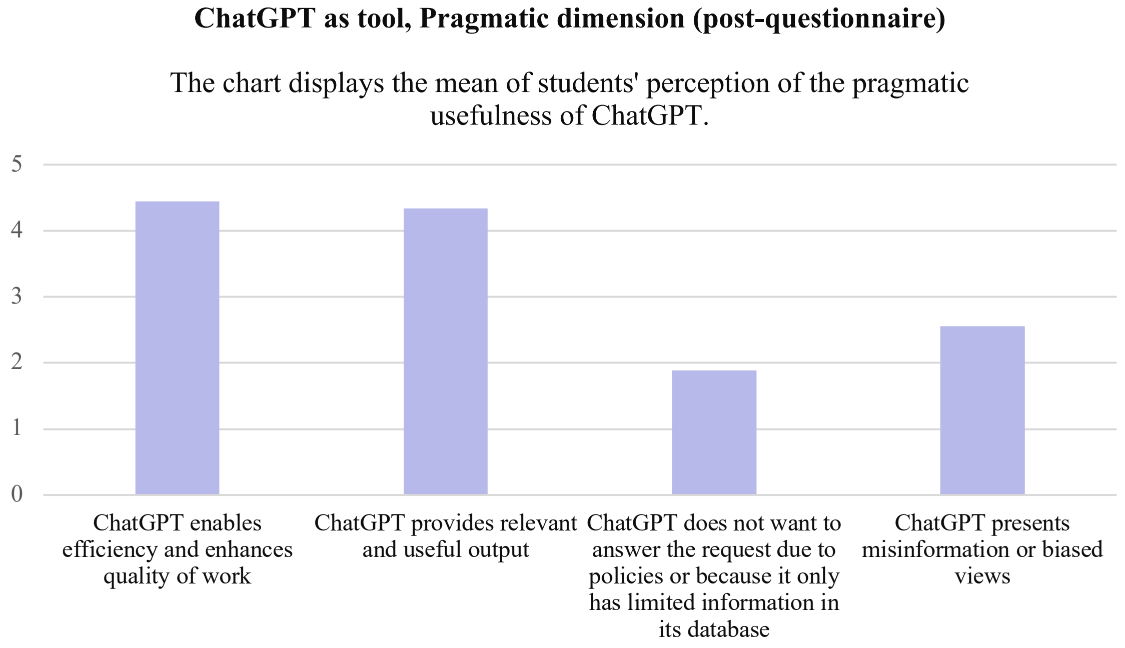


FIGURE 25*, ChatGPT as a tool, Pragmatic dimension.*

Analogously, the *Hedonic dimension*, i.e. the subjective, pleasurable or emotionally rewarding experience that the tool offers, exploring how fun, interesting the interactions are, also obtained positive results. Specifically, the results of the questions were: '*ChatGPT is exceeding expectations, impressive, or superior compared to existing solutions*' (m = 3.67, sd = 1.05), '*ChatGPT can support creative activities (such as essay writing, brainstorming or dialectical exchanges)*' (m = 4.33, sd = 0.67), '*Interactions with ChatGPT are entertaining*' (m = 3.89, sd = 0.99), '*Interactions with ChatGPT stimulate intellectual curiosity and engagement*' (m = 4.00, sd = 0.94). Findings indicate how ChatGPT is positively perceived both from a pragmatic point of view, considering it effective in accomplishing the promised tasks, and from a hedonic point of view, indicating it as a tool which not only fulfils but exceeds expectations, supporting creative activities and providing enjoyable and intellectually stimulating interactions. These outcomes suggest how it can be a valuable ally in the professional, academic and creative spheres.

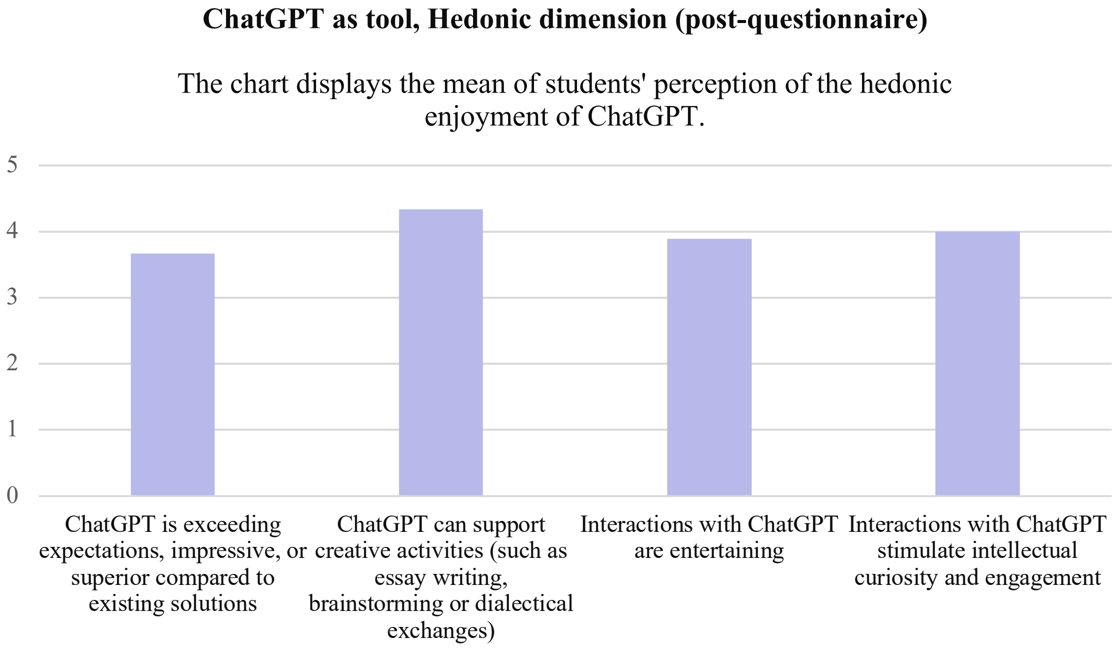


FIGURE 26*, ChatGPT as a tool, Hedonic dimension.*

As for the perceived *Human likeness* of the tool (Figure 27), on the other hand, the results were significantly lower, highlighting at the same time a critical awareness of the limitations exhibited by LLMs. Specifically, the results were: '*ChatGPT is humanlike or intelligent*' (m = 3.00, sd = 1.05), '*ChatGPT's responses seem to come from a real person*' (m = 2.56, sd = 1.17), '*ChatGPT exhibit human-like reasoning and comprehension*' (m = 3.00, sd = 0.94).

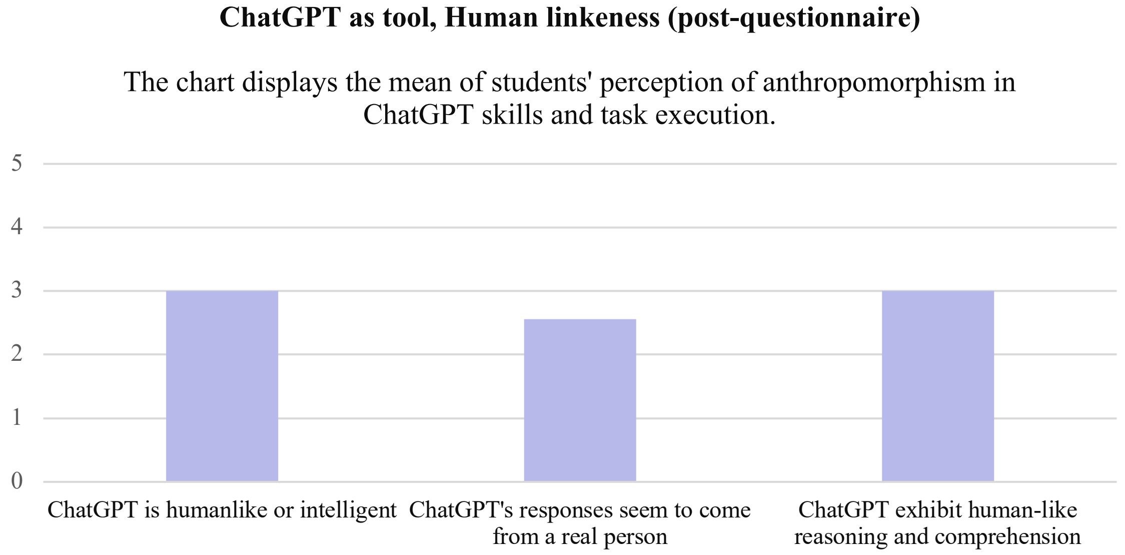


FIGURE 27*, ChatGPT as a tool, Human likeness.*

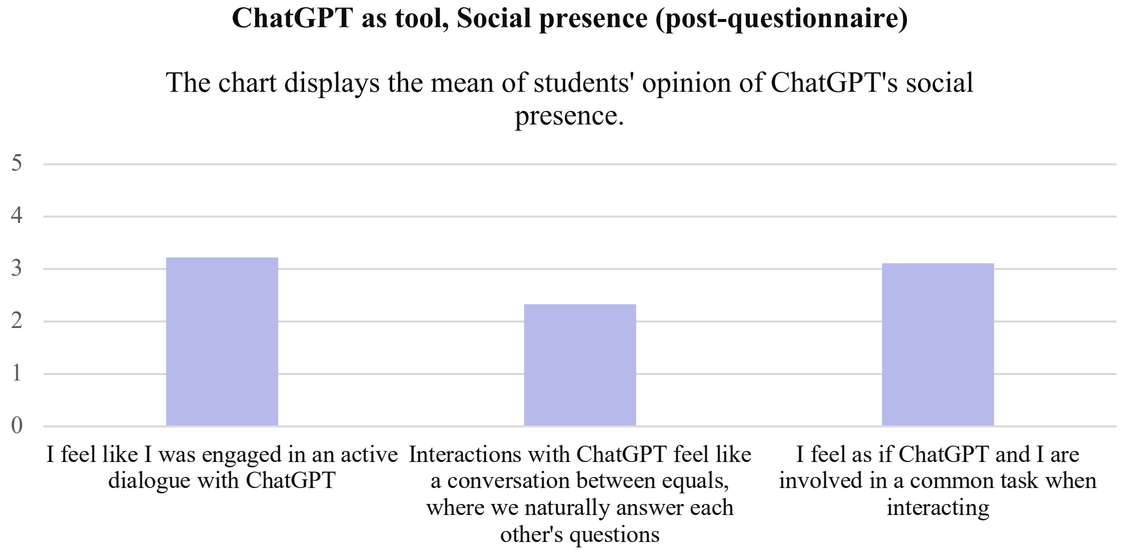


FIGURE 28*, ChatGPT as a tool, Social presence.*

The last section, *Social presence* (Figure 28), relates to the users' perception of being engaged in an active and meaningful dialogue with the chatbot, rather than simply interfacing with an automated system. Specifically, the results of the questions were: '*I feel like I was engaged in an active dialogue with ChatGPT*' (m = 3.22, sd = 0.79), ‘*Interactions with ChatGPT feel like a conversation between equals, where we naturally answer each other's questions*’ (m = 2.33, sd = 0.67), ‘*I feel as if ChatGPT and I are involved in a common task when interacting*’ (m = 3.11, sd = 0.74). In light of the findings, this section demonstrated participants’ broad degree of satisfaction with ChatGPT, emphasising its effectiveness in improving efficiency and offering useful and relevant answers, as well as its user-friendly interface to interact with. However, with respect to the perception of human likeness and social presence, the results show less enthusiastic evaluations: these imply that despite the answers may seem convincing, participants still perceive the difference between interacting with a human and an LLM. While users feel some cooperation in accomplishing tasks with ChatGPT, the programmatic nature of the interactions remains evident: these observations emphasise the significance of continuing to develop AI conversational technologies that can provide a more fluent experience, narrowing the gap in perceived naturalness and engagement in dialogue.

Perceived interaction improvement (post-questionnaire)

This is probably one of the most significant sections of the questionnaire, as it allowed students to express their feedback regarding the usefulness of the workshop held: in particular, specifically, the results were all highly positive: m = 4.52, sd = 0.57. The answers to the questions were: ‘*Since participating in the literacy course, I have noticed an improvement in my ability to interact effectively with ChatGPT*’ (m = 4.56, sd = 0.50), ‘*The literacy course has helped me better understand how to engage with ChatGPT for more meaningful interactions*’ (m = 4.56, sd = 0.50), ‘*After completing the literacy course, I feel easier to get useful responses from ChatGPT*’ (m = 4.44, sd = 0.68). These results indicate the effectiveness of this initiative on the participants‘ AIL: the positive feedback reflects the impact of the workshop on their digital competence, improving their ability to get useful responses and interact meaningfully with advanced technologies such as LLMs. This underlines the importance of integrating digital literacy education into curricula to enrich the educational experience and prepare students to face the challenges of the modern technological world.

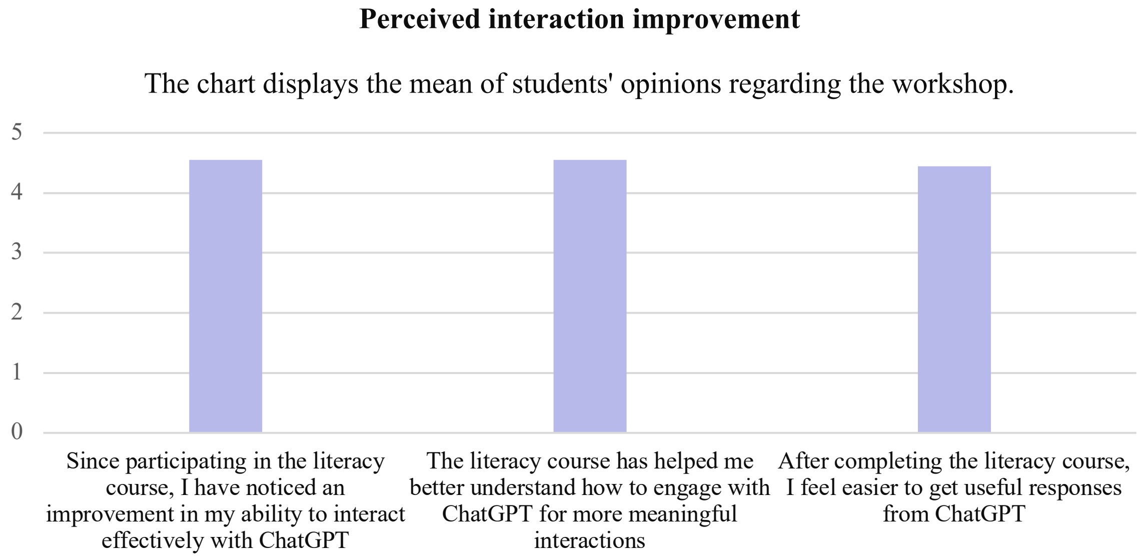


FIGURE 29*, Perceived interaction improvement.*

Ethical considerations (pre-questionnaire, post-questionnaire)

In the last section of the questionnaire, participants were asked for their opinions regarding the ethical implications of GAI and the future usefulness of AIL and prompt engineering, and in particular, following the workshop, these showed an improvement in the results: mPre = 4.20, sd = 0.79; mPost = 4.58, sd = 0.52. Specifically, in the *Ethical awareness* section, the question '*I am conscious of the potential biases embedded in large language models (LLMs) and their impact on user interactions*' (mPre = 4.20, sd = 0.79; mPost = 4.58, sd = 0.52) reported a p-value of 0.024 according to the Wilcoxon test, thus indicating a statistically significant change in terms of increased awareness of the biases of language models.

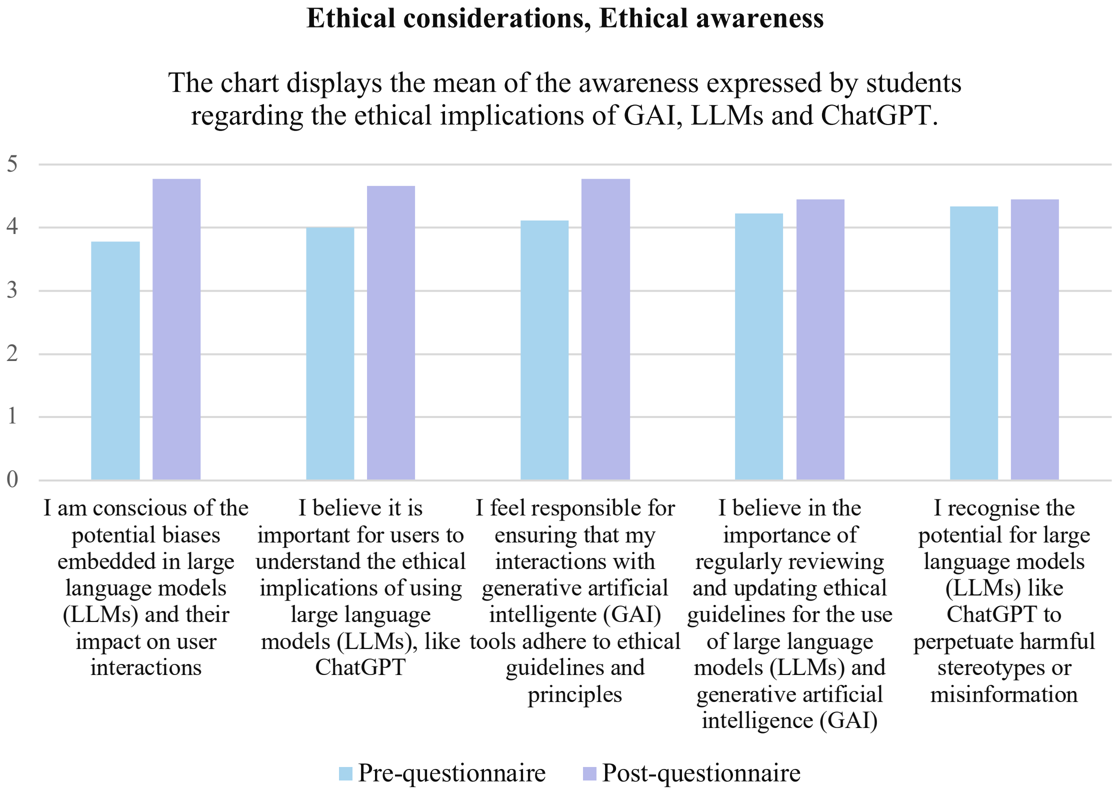


FIGURE 30*, Ethical considerations, Ethical awareness.*

Concurrently, also the question '*I believe it is important for users to understand the ethical implications of using large language models (LLMs), like ChatGPT*' (mPre = 4.00, sd = 0.94; mPost = 4.67, sd = 0.47) recorded a p-value of 0.059, if not inferior, relatively very close to the threshold, suggesting a shift in perception regarding the importance of spreading awareness in this area. The results of the other questions were: '*I feel responsible for ensuring that my interactions with generative artificial intelligence (GAI) tools adhere to ethical guidelines and principles*' (mPre = 4.11, sd = 0.87; mPost = 4.78, sd = 0.42), '*I believe in the importance of regularly reviewing and updating ethical* *guidelines for the use of large language models (LLMs) and generative artificial intelligence (GAI)*' (mPre = 4.22, sd = 0.63; mPost = 4.44, sd = 0.68), '*I recognise the potential for large language models (LLMs) like ChatGPT to perpetuate harmful stereotypes or misinformation*' (mPre = 4.33, sd = 0.68; mPost = 4.44, sd = 0.50).

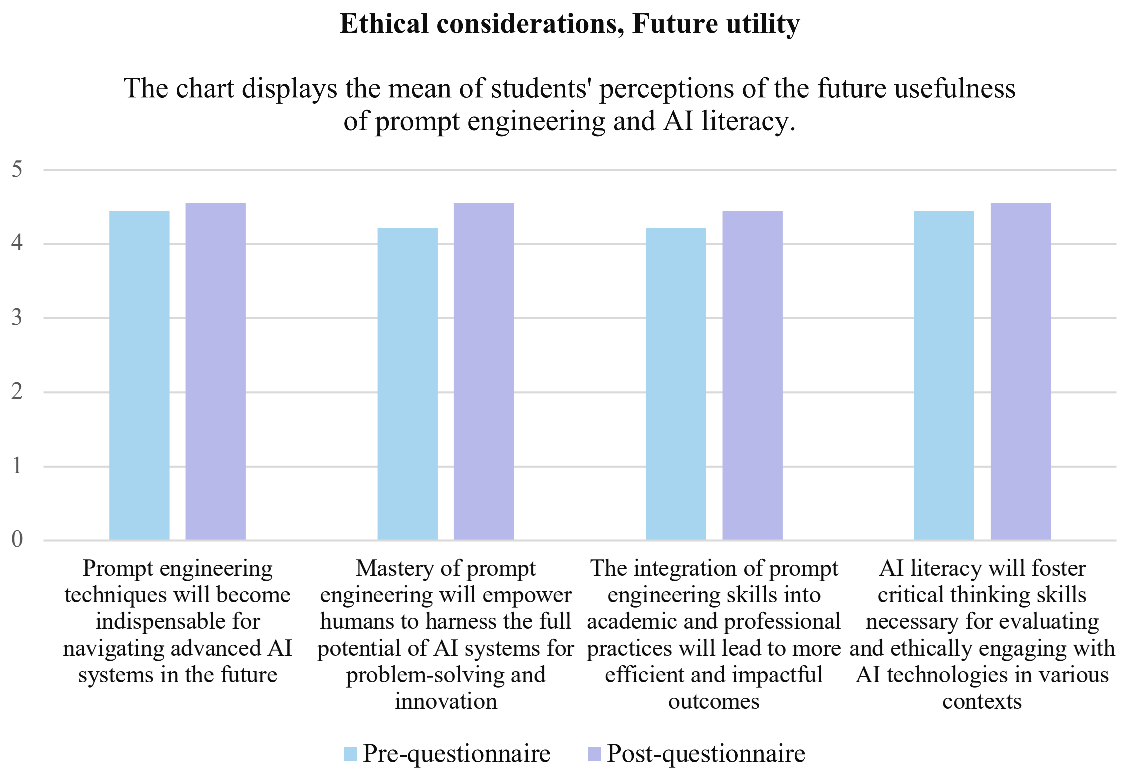


FIGURE 31, *Ethical considerations, Future utility.*

As for the *Future utility* section, this also showed an improvement. Specifically, The question '*Prompt engineering techniques will become indispensable for navigating advanced AI systems in the future*' (mPre = 4.44, sd = 0.68; mPost = 4.56, sd = 0.50) recorded a p-value of 0.046, statistically significant and therefore indicator of the effectiveness of the techniques presented according to the students. the results of the other questions were: '*Mastery of prompt engineering will empower humans to harness the full* *potential of AI systems for problem-solving and innovation*' (mPre = 4.22, sd = 0.63; mPost = 4.56, sd = 0.50), '*The integration of prompt engineering skills into academic and professional practices will lead to more efficient and impactful outcomes*' (mPre = 4.22, sd = 0.92; mPost = 4.44, sd = 0.50), '*AI literacy will foster critical thinking skills necessary for evaluating and ethically engaging with AI technologies in various contexts*' (mPre = 4.44, sd = 0.50; mPost = 4.56, sd = 0.50). The conclusions of this section therefore indicate a general enhancement in the ethical awareness and perceived importance of prompt engineering techniques as a consequence of the workshop. While the improvement is not always backed by statistical significance, there is evidence to suggest an increase in consciousness and understanding of the ethical implications associated with GAI models, particularly with regard to bias and the importance of wider awareness regarding these constraints. Concerning the future usefulness of artificial intelligence and prompt engineering techniques, the results indicate a perceived improvement, although again not statistically significant. Participants recognise the growing importance of these skills for the future, both in professional and academic settings, as well as their potential to facilitate innovation and problem-solving. This suggests that, despite the results not yet being sufficient to draw definitive conclusions, there is a positive trend.

Evaluation of the prompt engineering exercise

In terms of the evaluation of prompts and outputs, the results indicate a remarkable improvement in the interaction held during the post-questionnaire compared to the pre-questionnaire: mPre = 2.10, sd = 1.65; mPost = 3.89, sd = 1.57.

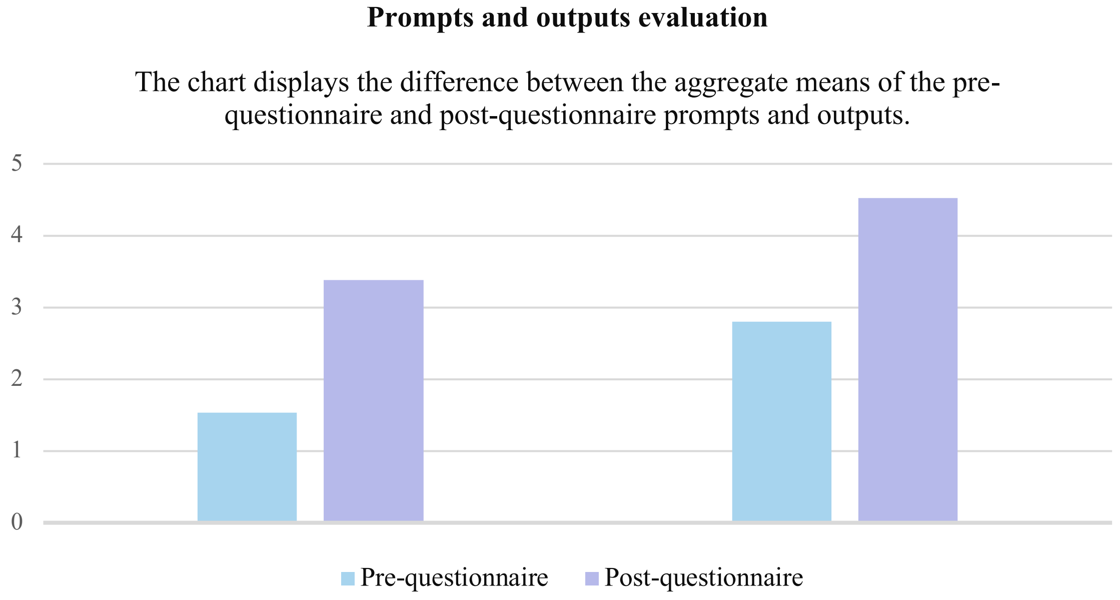


FIGURE 32, *Prompts and outputs evaluation.*

Analysing in detail the characteristics of the interactions, the improvement was measurable on several dimensions. Specifically, while in the pre-questionnaire the average number of pairs of inputs and outputs with which the students interacted with the model was 1 (sd = 0), in the post-questionnaire it rose to 1.89 (sd = 0.99), indicating greater interactivity in refining the responses obtained and more involvement in the conversation. Similarly, the mean of the number of words in the pre-questionnaire prompts was 65.1, while in the post-questionnaire it more than doubled, with a mean of 160.6 regarding the number of input words. However, only 22.22% of the students in the pre-questionnaire and 33.33% of the students in the post-questionnaire also provided an attachment to the prompt, i.e. the fictional curriculum vitae they had been given, indicating no substantial improvement in this particular regard.

When taking only prompts into consideration for the evaluation, assessed in accordance with the CLEAR framework, i.e. clear, logical, explicit, adaptive and reflective, which was suggested to the participants as a possible guidance for the formulation and iteration of effective prompts, a very significant improvement is noted: mPre = 1.53, sd = 1.36; mPost = 3.20, sd = 1.94. Through Spearman's correlation coefficient, also known as Spearman's rho (ρ), a statistical measure estimating the strength and direction of the monotonic relationship between two ordered variables, it was possible to verify how the number of input words is apparently weakly to moderately correlated with the ratings of the CLEAR framework categories. Specifically, for the '*Adaptive*' category of the pre-questionnaire, a correlation of 0.55 was verified: it is statistically moderately significant as it is close to ρ = 1, which instead denotes a perfect positive correlation (meaning that as the values of one variable increase, the values of the other variable also increase according to the exact order of the ranks). As for the post-questionnaire, the '*Clear*' and '*Explicit*' categories achieved a correlation with the prompt's word count of 0.63 and 0.67 respectively, indicating a moderately positive correlation and close to being statistically significant. This implies that overall, the number of input words does not consistently influence the evaluation across the measured categories, other than moderately in some categories. Nevertheless, the evidence regarding the improvement in the formulation of prompts is statistically significant for almost all categories. Specifically, '*Clear*' (mPre = 2.89, sd = 0.57; mPost = 4.33, sd = 0.94) and '*Explicit*' (mPre = 2.00, sd = 0.94; mPost = 4.33, sd = 1.25), with the Wilcoxon test detecting for both a p-value of 0.010, which is therefore statistically significant. Furthermore, the '*Adaptive'* (mPre = 0.11, sd = 0.31; mPost = 2.00, sd = 2.11) and '*Reflective*' (mPre = 0.00, sd = 0.00; mPost = 2.11, sd = 2.38) categories, reached p-values of 0.041 and 0.034 respectively, indicating a positive improvement trend. On the other hand, the ‘*Logic’* category (mPre = 2.67, sd = 0.47; mPost = 3.22, sd = 1.03) was not statistically significant according to the Wilcoxon test, but still improved.

Evaluation of prompts

The chart displays the mean of the features evaluated through the CLEAR framework of the prompts from the pre and post-questionnaire practical test.

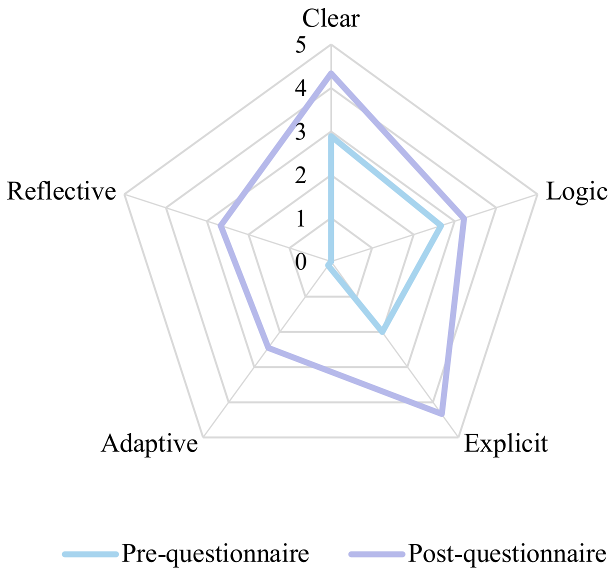


FIGURE 33*, Evaluation of prompts.*

Delving further into the examination of the prompts, it is notable that many of those in the practice test regarding the pre-questionnaire contained expressions such as «Hi», «Hey» or «Please», as they would often be included in a conversation between humans, indicating an early influence of the ELIZA effect in terms of politeness of speech. Furthermore, the prompts in the first trials were generally generic and less detailed, for instance «Create a motivational letter, in British English, for a PISE student for the University College of London», without any details to orient the model as to the specifics of the assignment, and assuming information such as the significance of each letter within the acronym of the faculty, or «Please, could you write a motivational letter for a MA in Digital Ethics at UCL?». Remarkably, many of these expressions were not repeated in the post-questionnaire exercises. The students were more accurate in providing details and additional background information, as well as more critical in evaluating the output received: in particular, by quoting a portion of the motivational letter generated by ChatGPT in the first output received, one student specified «Not something I told u», and later also «Cut this part as I haven't given u info about what I want to do with my degree», whereas another student in his/her own interaction submitted the prompt «I don't like the part where [you made] a personal connection to the growing importance of digital ethics, explaining how [I] hope to contribute to the field both during and after the course... I prefer to explain what positive contribution I want to spark with my ideas, ultimately contributing to EU's digital ethics framework». Such efforts signal a greater awareness and attention to the errors made by the tool, and a more iterative approach as encouraged during the workshop. Similarly, the request «But do not invent information: if you do not know something, ask a few questions until you have every clear piece of information to write a coherent letter», demonstrated an awareness and knowledge of the inherent limitations of LLMs and their ability to generate erroneous and misaligned outputs. An interaction also took a reverse approach: specifically, in order to avoid hallucinations, the student prompted ChatGPT to «Ask me as many questions as you can so that I know the information you need to write a super-detailed, hallucination-free, personal and motivated letter». Therefore, the students effectively proved to realistically have internalised the notions of the workshop and applied them in real-world contexts.

Regarding the outputs, these also witnessed a significant improvement as a result of the AIL workshop and the prompt engineering and prompt patterns techniques: mPre

= 2.81, sd = 1.72; mPost = 4.53, sd = 0.69. Specifically, the category experiencing a marked improvement was that of '*Originality*' (mPre = 2.22, sd = 0.79; mPost = 4.44, sd = 0.50), with a p-value of 0.004, thus suggesting a significant difference of the originality of the outputs produced in the post-questionnaire compared to the pre-questionnaire.

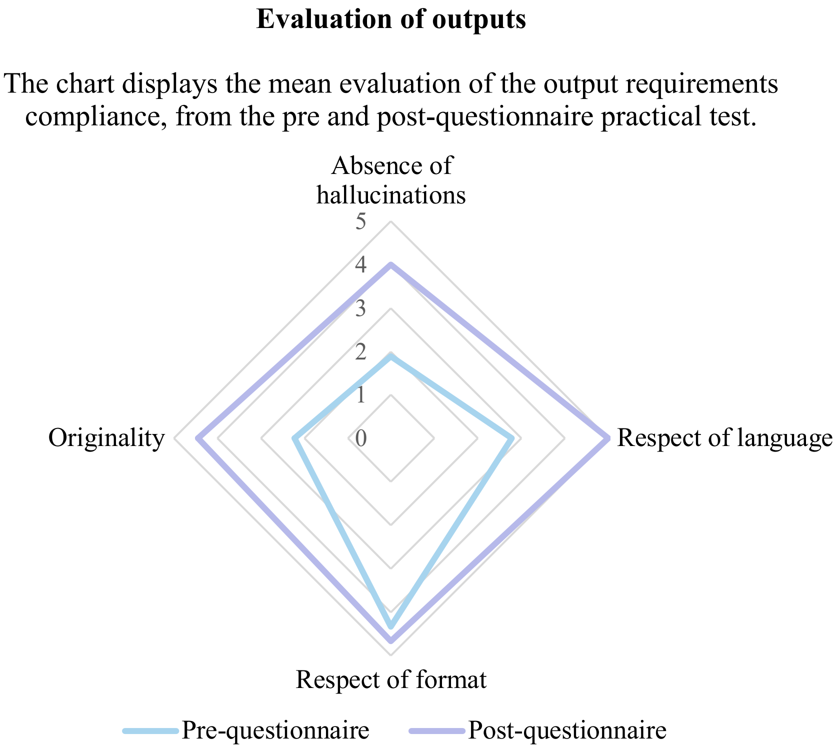


FIGURE 34*, Evaluation of outputs.*

In parallel, the categories '*Absence of hallucinations*' (mPre = 2.22, sd = 0.79; mPost = 4.44, sd = 0.50) and '*Respect of language*' (mPre = 2.22, sd = 0.79; mPost = 4.44, sd = 0.50) obtained p-values of 0.017 and 0.046 respectively, revealing a statistically significant improvement. Conversely, the category '*Respect of format*' (mPre = 4.33 sd = 0.67; mPost = 4.67, sd = 0.47), while seeing a positive trend, did not achieve a statistically significant p-value, thus indicating a minor change.

Therefore, taking into account all the evidence discussed so far, it is possible to highlight a significant improvement in the quality of students' interactions with ChatGPT between the pre and post-questionnaire, revealing an advancement in their ability to formulate more articulate and focused prompts. The greater interactivity, measured through the number of input-output pairs and length of prompts, suggests increased engagement, reflecting a heightened recognition of the importance of refining prompts and critically evaluating the model's responses. Improvements in prompts did not manifest consistently across all categories of the CLEAR framework: the categories related to clarity and explicitness showed statistically significant improvement, however the categories related to adaptability and reflexivity, although having shown improvement, only approximated a threshold of statistical significance. This could suggest that the development of these skills demands a more in-depth and extended training approach to consolidate effectively. In terms of outputs, the results are also indicative of an improvement in overall quality, with notable progress in the categories related to originality and absence of hallucinations. This implies at the same time that the introduction of prompt engineering techniques contributed to greater accuracy and creativity in the responses produced. On the whole, these findings show a positive and measurable impact of the workshop on the quality of the interactions and outputs generated by the students.

VII.1. Discussion of findings

The pilot study carried out allowed for an in-depth exploration of the impact of literacy in artificial intelligence and prompt engineering techniques on a group of humanities (non-STEM) university students. The purpose of the workshop was to improve the participants' understanding of the functioning, ethical and legal implications and practical handling of GAI tools, such as LLMs. Despite the participants' limited initial knowledge of AI ('*Limited knowledge*' 55.56%) and GAI ('*Limited knowledge*' 66.67%), and hardly any knowledge of prompt engineering ('*No knowledg*e' 55.56%), the workshop showed significant positive effects on several fronts, indicating the potential of such educational interventions in the academic context and beyond. Specifically, ChatGPT, the chatbot developed by OpenAI, was used, with the aim of deepening university students' understanding of the limitations of AI and the effectiveness of their interactions with this system. As verified by the questionnaire, the majority of students (66.67%) use it '*Very often*', not only for personal but also for academic purposes.

In the first place, a decrease in the perceived realistic and identity threat associated with AI was observed after attending the workshop (from a mean of 3.79 to 2.56). This change suggests that appropriate AIL, allowing for an understanding of the capabilities, functioning, and inherent limitations of GAI, as well as guidance experiences can effectively mitigate and modulate concerns associated with these technologies. Specifically, a reduction in concerns regarding job loss and the belief that AI is blurring the boundaries between humans and machines was noted. The question '*The realism of artificial intelligence is disturbing because it makes it almost indistinguishable from* *human beings*' recorded a change in mean from 3.33 to 2.44, with a p-value of 0.054, very close to the threshold of 0.05. Nevertheless, it is important to note that some deep-seated concerns, such as those regarding the challenge to the essence of human beings posed by AI, showed a less marked reduction, and therefore persist and demand further educational interventions and ongoing dialogue.

In terms of emotional experience, students reported predominantly positive emotions during interaction with ChatGPT, such as '*Serenity*' (mean of 3.78) and '*Joy*' (mean of 3.33), as opposed to less intense negative emotions such as '*Anger*', '*Fear*' and '*Disgust*'. This points not only to a favourable acceptance of the technology, but also to an enthusiasm towards its usage, potentially facilitating the integration of GAI tools into every day and academic practices. It remains crucial, however, to monitor and understand how these emotions may influence users' critical judgement of emerging technologies. This might be a fertile ground for further structured investigations and empirical tests.

The post-workshop data also show a slight overall improvement in the quality assessment of ChatGPT interactions (from a mean of 3.17 to 3.19), however with a large improvement in ChatGPT's comprehension and response capabilities. The question '*I am satisfied with ChatGPT's comprehension and response capabilities*', with a mean of 3.33 before and 4.00 after, recorded a p-value of 0.014, thus indicating a statistically significant improvement. This demonstrates the effectiveness of the prompt engineering and prompt patterns techniques introduced during the workshop. The question '*ChatGPT repeats the same mistakes over and over again, without adapting to my questions*' also showed a modest improvement, from a mean of 2.78 to a mean of 2.33. With regard to the effort required to obtain the desired output from the chatbot, there was a significant increase with respect to the question '*I found it easy to communicate my intentions to ChatGPT*', with a mean of 3.44 and following the workshop of 4.11, while suggesting how a structured approach to prompt formulation can help students in their interactions with machines.

In terms of the evaluation of ChatGPT as a tool, the majority of the participants showed a positive opinion both from the point of view of the pragmatic dimension and the hedonistic dimension. In particular, the question '*ChatGPT enables efficiency and enhances quality of work*' scored a mean of 4.44, underlining how it is perceived as a facilitator and enhancer of human intelligence in professional, academic and creative processes. Similarly, the question '*ChatGPT can support creative activities (such as essay writing, brainstorming or dialectical exchanges)*' achieved a mean of 4.33. The perception of human likeness and social presence of the tool was less pronounced, revealing both an underlying limitation of the system in the naturalness of the interactions it enables, but at the same time a solid awareness of the students regarding the technology's limitations and the substantial differences from human cognitive processes, a topic that was particularly emphasised during the workshop.

Participants showed very enthusiastic opinions regarding the effectiveness of the workshop conducted, with an average of 4.52. This result is of key importance in assessing the impact and success of the design of the training session in terms of structure and information selection, a crucial result if it is to be replicated on a larger scale. Notably, after participating in the workshop, students reported an improvement in their ability to interact with ChatGPT, while at the same time finding it easier to obtain useful answers from the tool. The responses highlight the importance of integrating this type of educational activity into curricula to prepare individuals for the challenges of the contemporary technological world, as well as to ensure they can exploit the tool's functionalities to the full (in terms of personalised learning, for instance).

The AIL also led to an enhancement in ethical awareness, from a mean of 4.20 to 4.58. Specifically, there was an increase in awareness concerning the existence of biases in LLMs: the question '*I am conscious of the potential biases embedded in large language models (LLMs) and their impact on user interactions*' registered a mean of 4.20 and 4.58, before and after the workshop respectively, with a p-value recorded by Wilcoxon's test of 0.024, thus indicating a statistically significant change. The question '*I believe it is important for users to understand the ethical implications of using large language models (LLMs), like ChatGPT*' also recorded a relevant p-value of 0.059. Furthermore, the students also indicated an improved responsibility to ensure that their interactions adhere to ethical guidelines. These findings are meaningful, as they underline the importance of including ethics in any AI literacy journey, while at the same time pointing to the usefulness of raising awareness about these matters, frequently seen as secondary to the technical capabilities and evident limitations of LLMs and GAI. Participants also recognised the significance of prompt engineering in navigating more consciously in an AI-enhanced future, both academically and professionally, as well as the importance of AI literacy in providing the critical thinking skills necessary to interact effectively and responsibly with these technologies.

On a final note, the prompt engineering assignment also showed significant improvements, both in terms of the formulation of the prompts, with a mean rating according to the CLEAR framework of 1.53 before the workshop and 3.20 thereafter, as well as in terms of the results obtained and adherence to the directions provided by the exercise delivery, with a mean of 2.81 and subsequently 4.53. In addition to that, there was over a doubling with respect to the average number of words included in the prompts, which prior to the workshop was 65.1 and later became 150.3, along with the input-outputs pairs increasing from a mean of 1.00 to 1.67, indicating that students spent more time and developed more detailed instructions to guide the LLM, displaying deeper engagement and advanced understanding of the importance of providing rich contexts. Notably, when analysing the specific categories of the CLEAR framework, the '*Clear*' and '*Explicit*' showed relevant progress: the average score for '*Clear*' increased from a mean of 2.89 to 4.33, and for '*Explicit*' rose from 2.00 to 4.33, both with p-values of 0.010, therefore statistically significant according to the Wilcoxon test. Also the '*Adaptive'* (mPre = 0.11, sd = 0.31; mPost = 2.00, sd = 2.11) and '*Reflective*' (mPre = 0.00, sd = 0.00; mPost = 2.11, sd = 2.38) categories, reached p-values of 0.041 and 0.034 respectively, indicating a positive improvement trend. This indicates that the students developed an initial ability to adapt and iterate their prompts according to the outputs obtained as well as to reflect on the interaction process, competences that could be further strengthened with additional training. In terms of the generated outputs, there was a significant improvement in the overall mean score, from an average of 2.81 to 4.53. Specifically, the '*Originality*' category witnessed an increase of its mean from 2.22 to 4.44, with a p-value of 0.004, showing how the participants were able to elicit more creative and less standardised responses from the model. The '*Absence of hallucinations*' and '*Respect of language*' categories also showed considerable enhancements, hinting at greater accuracy and relevance of the responses produced. These findings underline the effectiveness of the workshop in improving students' prompt engineering skills. The ability to formulate clear and explicit prompts led to more fruitful interactions with

ChatGPT, while the increase in originality and accuracy in the outputs reflects a deeper understanding of how to guide the GAI towards desired outcomes.

As with any other study, there are limitations to this one which must be acknowledged, as they could affect the interpretation of data and its more widespread applicability. One major constraint relates to the limited size of the sample: as previously specified, the study involved only 9 students, a number which, while providing useful preliminary data, may not be sufficient to draw generalisable conclusions. With such a relatively narrow sample, the results may reflect specific characteristics of this group rather than indicating valid tendencies for a larger population of university students: consequently, in a future study, it would be desirable to enlarge the sample in order to obtain more robust and representative outcomes. Another significant limitation stems from the voluntary nature of participation: the students self-selected, and this introduces a possible bias, as the individuals who chose to participate might be likely to be already more interested in technology and AI or GAI, biasing their perceptions towards ChatGPT and prompt engineering techniques. Furthermore, the emotions and opinions expressed by the participants could be affected by different temporary factors or initial enthusiasm towards the utilisation of the technological tools, without fully reflecting the actual long-term benefits. This might have resulted in an overestimation of the observed improvements, as these participants could be more predisposed to interact positively with AI tools than the average humanities student. In addition, while this group was selected precisely because their backgrounds may be less technical, the findings might not be applicable to students from other disciplines, such as scientific or technological fields, who are likely to have different experiences and needs when interacting with GAI systems and LLMs. Ultimately, the responses to the questionnaires and the evaluation of prompts and outputs, which formed the foundation for the workshop evaluation, reflect subjective student and evaluator perceptions and may not fully capture the objective effectiveness of the AIL workshop and prompt engineering techniques. These limitations suggest the need for further research to corroborate and extend the findings, with larger samples, less biased recruitment methodologies, as well as longer training programmes to address more subjects, more examples, and more tailored and advanced prompt engineering techniques.

In conclusion, the pilot study demonstrated how a literacy workshop on artificial intelligence, generative artificial intelligence and prompt engineering can be successful for humanities students with limited initial knowledge of AI can improve their knowledge, awareness and interactions with these technologies. Engaging with undergraduate students from the humanities, it highlighted the specific needs of an audience which is often less exposed to advanced technologies, highlighting gaps in knowledge and misperceptions about artificial intelligence and LLMs such as ChatGPT. This underscored the importance of developing AIL training programmes, regardless of students' backgrounds. The workshop, structured in theoretical and practical training modules, proved the effectiveness of an integrated approach to AIL; as the combination of comprehensive explanations on how LLMs work with practical exercises on prompt engineering fostered active learning, allowing students to immediately apply the acquired knowledge and improve their ability to interact effectively with GAI tools. Furthermore, the study emphasised the importance of including discussions on the ethical, legal and social implications of AI in educational approaches, displaying a significant increase in students' awareness of the inherent biases in LLMs and the importance of ethical and responsible use of these technologies. Analysis of the pre- and post-workshop questionnaires yielded empirical evidence on the positive impact of AIL on students' skills, with significant improvements in the formulation of prompts, quality of interactions with AI, and perceptions of the potential and limitations of GAI models. These data can serve to guide the design of effective educational curricula: building on the findings, it is possible to propose a framework for AI literacy in higher education that includes an initial assessment of students' knowledge, theoretical modules adapted to the disciplinary context, practical exercises on prompt engineering, ethical discussions, and ongoing assessment to monitor learning. The outcome of the pilot study suggests that this model can be adapted to a broader audience and to different academic disciplines, facilitating a wider dissemination of these initiatives in Italian universities. By embracing AIL, in fact, academic institutions can prepare students to face the challenges posed by the increasing integration of AI in society, providing them not only with technical skills, but also with the ability to critically consider and use technology in an ethical and responsible manner.

CONCLUSIONS

This thesis thoroughly researched the importance of AI literacy in the context of technological evolution and its impact on society, with a specific focus on GAI tools – such as LLMs and systems like ChatGPT. This work provided an articulate treatment not only from a theoretical point of view, by analysing the historical, conceptual and ethical development of generative artificial intelligence, but also through a pilot study aimed at exploring how AIL and prompt engineering can improve the knowledge, awareness, and practical and critical skills of non-STEM university students.

The first part of the thesis focused on a theoretical and historical overview on AI, GAI and HAII, defining the key concepts and the evolution. The importance of embracing a human-centred approach was emphasised, not only in the development of these systems, but also in the beneficial collaboration between AI and human intelligence, ensuring that this technology is an augmentation (and not a replacement) of people's decision-making faculties. In this respect, prompt engineering and prompt patterns can help users communicate effectively with LLMs, resulting in more coherent and aligned outputs. Ethics and critical reflection have also formed a very important thread throughout this thesis. It is essential to adopt a responsible and ethical attitude, respecting human rights and fostering fairness, transparency and sustainability in all aspects of GAI.

Furthermore, the concept of artificial intelligence literacy was one of the main pillars of the thesis. It was interpreted as a set of transversal knowledge and competences enabling critical, collaborative and ethical interaction with GAI systems. This literacy is not restricted to the technical understanding of machine learning models or neural networks, but encompasses the ability to critically evaluate the limitations, risks and ethical implications of using these tools. The need to educate for ethical awareness emerged strongly throughout the work. In an age where GAI forms an integral part of operational and creative processes at every level, frequently with significant decision-making consequences, it is paramount for individuals to be equipped with the necessary means to comprehend the full potentials and consequences of these technologies. This includes the ability to recognise bias in AI models, possible privacy violations, and the impact such systems may have on society as a whole, in terms of equity and justice. This includes the ability to recognise bias in AI models, possible privacy violations, and the impact such systems may have on society as a whole, in terms of equity and justice. The chapter dedicated to education (chapter IV) examined in detail the initiatives underway, both at European and Italian national level, to promote AIL, from primary school to university. Italian universities, while recording some progress, were reported as still far from creating a cohesive and inclusive framework that would guarantee all students, regardless of their academic background, the opportunity to develop AI and prompt engineering skills.

The second part of the thesis presented the empirical study, which formed an innovative component of this work. The workshop organised at the Ca’ Foscari University of Venice, aimed at a sample of humanities students, sought to bridge the AI-related education gap among students from non-STEM backgrounds. The pilot demonstrated how a structured AI literacy programme can produce a positive and measurable impact on the ability to interact with LLMs-based chatbots such as ChatGPT. Specifically, the participants demonstrated an improvement in their prompt engineering skills, their ability to understand the inherent limitations of LLMs, as well as an increased critical awareness of the ethical use of such tools. The most remarkable aspect of the study was the significant enhancement in prompt engineering skills, both in qualitative and quantitative terms. Students displayed an increasing ability to articulate more detailed requests, to iterate and refine prompts, and to interact more thoroughly with the model, thus raising the quality of the responses obtained. From an ethical perspective, the study showed an increase in participants' awareness of the potential biases embedded in LLMs and the need to ensure responsible and informed use of AI and GAI. This ethical awareness, together with the technical expertise gained, is crucial to ensuring that artificial intelligence is used in a fair, transparent and responsible manner.

In conclusion, this thesis demonstrated how well-structured and integrated AIL in education is important for developing a critical and informed understanding of AI technologies. Through the theoretical analysis of how AI works, its limitations, challenges and opportunities, prompt engineering techniques, and a literature review regarding AIL and the educational landscape, as well as through the implementation of a focused empirical study, it was possible to explore how AI literacy skills can positively influence university students’ ability to interact with ChatGPT.

Genuine progress lies not only in technology, but in the ability to employ it wisely and responsibly. Therefore, the future does not depend on machines, but to those who will be able to comprehend them and interact with them ethically and intelligently. *Cogito, ergo sum.*

REFERENCES

GENERAL REFERENCES

ABEDIN, BABAK, ET AL., *Designing and Managing Human-AI Interaction*, in «Information System Frontiers», vol. 24, 2022, pp. 691-697, https://doi.org/10.1007/s10796-022-10313-1, accessed on 23.08.2024.

AHDADOU, MANAL, AAJLY, ABDELLAH, TAHROUCH, MOHAMED, *Unlocking the potential of augmented intelligence: a discussion on its role in boardroom decision-making*, in «International Journal of Disclosure and Governance», vol. 21, 2024, pp. 433-446, https://doi.org/10.1057/s41310-023-00207-2, accessed on 03.09..2024.

AHMED, AHMED AWAD E., *Employee Surveillance Based on Free Text Detection of Keystroke Dynamics*, in MANISH GUPTA, RAJ SHARMAN, *Handbook of Research on Social and Organizational Liabilities in Information Security*, Hershey, IGI Global, 2009.

ARISTOTLE, *Organon*, edited by Giorgio Colli, Adelphi Edizioni, Milan, 2003.

AVERBUKH, VLADIMIR L., *Sources of Computer Metaphors for Visualization and Human-Computer Interaction*, in *IntechOpen*, 17 Jun. 2020, https://doi.org/10.5772/intechopen.89973, accessed on 01.09.2024.

AUERNHAMMER, JAN , *Human-centered AI: The role of Human-centered Design Research in the development of AI*, in STELLA BOESS, MING CHEUNG, REBECCA CAIN (edited by), «Proceedings of DRS2020 International Conference», 2020, https://doi.org/10.21606/drs.2020.282, accessed on 01.09.2024.

BASSET, CAROLINE, *The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present*, in «AI & Society», volume 34, pp. 803-812.

BERRY, DAVID M., *The Limits of Computation*, in «Weizenbaum Journal of the Digital Society» vol. 3, n. 3, 2023, https://doi.org/10.34669/WI.WJDS/3.3.2, accessed on 23.08.2024.

BOINE, CLAIRE, *Emotional Attachment to AI Companions and European Law*, in *MIT Case Studies in Social and Ethical Responsibilities of Computing*, 27 Feb. 2023, https://doi.org/10.21428/2c646de5.db67ec7f, accessed on 02.09.2024.

CHEN, MARK, ET AL., *Generative Pretraining from Pixels*, in *OpenAI*, https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf, accessed on 19.07.2024.

CHESTERMAN, SIMON, *Good models borrow, great models steal: intellectual property rights and generative AI*, in «Policy and Society», 2024, https://doi.org/10.1093/polsoc/puae006, accessed on 20.08.2024.

CHOMSKY, NOAM, *Syntactic structures*, Berlin, New York, De Gruyter Mouton, 2002.

CHRISTIE'S, *Obvious and the interface between art and artificial intelligence*, 12 Dec. 2018, https://www.christies.com/en/stories/a-collaboration-between-two-artists-one-human-one-a-machine-0cd01f4e232f4279a525a446d60d4cd1, accessed on 20.08.2024.

CODENOTTI, BRUNO, LEONCINI, MAURO, *Apprendimento Automatico*, in IID., *La Rivoluzione silenziosa: le grandi idee dell'informatica alla base dell'era digitale*, Turin, 2020, pp. 127-156.

COLDEWEY, DEVIN, *OpenAI shifts from nonprofit to 'capped-profit' to attract capital*, in «TechCrucnh», 11 Mar. 2019, https://techcrunch.com/2019/03/11/openai-shifts-from-nonprofit-to-capped-profit-to-attract-capital/, accessed on 03.09.2024.

CONLON, ED, DABUS: South Africa issues first-ever patent with AI inventor, in «Managing IP», 29 Jul. 2021, https://www.managingip.com/article/2a5czh91g6c8zwxjcpla8/dabus-south-africa-issues-first-ever-patent-with-ai-inventor, accessed on 20.08.2024.

COURT OF JUSTICE OF THE EUROPEAN UNION, Judgment of 16 July 2009, *Infopaq International A/S v Danske Dagblades Forening, Case C-5/08*, paragraphs 46-47.

CRISTIANINI, NELLO, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*, Il Mulino, Bologna, 2023.

DAVIS, MARTIN, *The Universal Computer*, New York, Norton, 2000.

DE SAUSSURE, FERDINAND, *Cours de linguistique générale*, introduction, translation and commentary by Tullio de Mauro, Editori Laterza, Bari, 1967.

DÍAZ-RODRÍGUEZ, NATALIA, et al., *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*, in «Information Fusion», vol. 99, 2023, https://doi.org/10.1016/j.inffus.2023.101896, accessed on 20.08.2024.

DIX, ALAN, *Human-Computer Interaction*, in LING LIU, M. TAMER ÖZSU, *Encyclopedia of Database Systems*, Boston, Springer, 2009, https://doi.org/10.1007/978-0-387-39940-9_192, accessed on 20.08.2024.

EUROPEAN COMMISSION, *Ethics Guidelines for Trustworthy AI*, 8 Apr. 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, accessed on 20.05.2024.

FLORIDI, LUCIANO, ET AL., *How to Design AI for Social Good: Seven Essential Factors*, in ID., *Ethics, Governance, and Policies in Artificial Intelligence*, Cham, Springer, 2021.

GHOSH, SHIKHAR, BAGAI, SHWETA, *Anthropic: Building Safe AI*, in «Harvard Business School Case 824-129», 2024.

HARMON-JONES, CINDY, BASTIAN, BROCK, HARMON-JONES, EDDIE, *The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions*, in «PloS one», vol. 11, n. 8, 2016, https://doi.org/10.1371%2Fjournal.pone.0159915, accessed on 07.04.2024.

HOFSTADTER, DOUGLAS RICHARD, FLUID ANALOGIES RESEARCH GROUP, *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanism of Thought*, Basic Books, New York, 1995.

HUANG, KALLEY, *Why Pope Francis Is the Star of A.I.-Generated Photos*, in «The New York Times», 08 Apr. 2023, https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html, accessed on 15.06.2024.

JOUPPI, NORMAN PAUL, ET AL., *In-Datacenter Performance Analysis of a Tensor Processing Unit*, paper presented at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, 2017, https://doi.org/10.48550/arXiv.1704.04760, accessed on 20.07.2024.

JUDGE, BRIAN, NITZBERG, MARK, RUSSEL, STUART, *When code isn't law: rethinking regulation for artificial intelligence*, in «Policy and Society», 2024, https://doi.org/10.1093/polsoc/puae020, accessed on 20.07.2024.

KARPATHY, ANDREJ, «The hottest new programming language is English», in *X*, https://x.com/karpathy/status/1617979122625712128?lang=en, accessed on 02.04.2024.

KHAN, ARIF ALI, ET AL., *Ethics of AI: A Systematic Literature Review of Principles and Challenges*, in *ArXiv*, https://doi.org/10.48550/arXiv.2109.07906, accessed on 13.09.2024.

MCCARTHY, JOHN, ET AL., *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 31 Aug. 1955, https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, accessed on 20.02.2024.

MICKLE, TRIPP, RENNISON, JOE, *Nvidia Becomes Most Valuable Public Company, Topping Microsoft*, in in «The New York Times», 18 Jun. 2024, https://www.nytimes.com/2024/06/18/technology/nvidia-most-valuable-company.html, accessed on 20.07.2024.

MITCHELL, MELANIE, *Artificial intelligence: a guide for thinking humans*, London, Pelican, 2019.

LEIBNIZ, GOTTFRIED WILHELM, *Leibniz: Selections,* translated by Philip Paul Wiener, New York, Charles Scribner's Sons, 1951.

LIGHTHILL, JAMES, *Artificial Intelligence: A General Survey*, in «Artificial Intelligence: a paper symposium», London, Science Research Council, 1973.

MANN, SEBASTIAN PORSDAM, ET AL., *AUTOGEN and the Ethics of Co-Creation with Personalized LLMs – Reply to the Commentaries*, in «The American Journal of Bioethics», vol. 24, n. 3, 2024, https://doi.org/10.1080/15265161.2024.2308175, accessed on 20.08.2024.

—, ET AL., *Generative AI entails a credit–blame asymmetry*, in «Nature Machine Intelligence», vol. 5, 2023, pp. 472-475, https://doi.org/10.1038/s42256-023-00653-1, accessed on 20.08.2024.

METHNANI, LEILA, ET AL., *Let Me Take Over: Variable Autonomy for Meaningful Human Control*, in «Frontiers Artificial Intelligence», vol. 4, 2021, https://doi.org/10.3389/frai.2021.737072, accessed on 24.07.2024.

MINSKY, MARVIN LEE, PAUPERT, SEYMOUR AUBREY, *Perceptrons: an Introduction to Computational Geometry*, Cambridge (Massachusetts), The MIT Press, 1969.

NORDEMANN, JAN BERND, PUKAS, JONATHAN, *Copyright exceptions for AI training data—will there be an international level playing field?*, in «Journal of Intellectual Property Law & Practice», vol. 17, 2022, pp. 973-974, https://doi.org/10.1093/jiplp/jpac106, accessed on 20.03.2024.

NORMAN, DONALD, *The Design of Everyday Things*, Cambridge, MA, The MIT Press, 2013.

NOVELLI, CLAUDIO, ET AL., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurit*y, in *ArXiv*, 15 Mar. 2024, https://doi.org/10.48550/arXiv.2401.07348, accessed on 20.08.2024.

OVALLE, ANAELIA, ET AL., *Queer In AI: A Case Study in Community-Led Participatory AI*, in «FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency», pp. 1882-1995, https://doi.org/10.1145/3593013.3594134, accessed on 20.06.2024.

ROSENBLATT, FRANK, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*, in «Psychological Review», volume 65, number 6, 1950.

RUMELHART, DAVID EMERETT, MCCLELLAND, JAMES LLOYD, PDP RESEARCH GROUP, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, The MIT Press, Cambridge (Massachusetts), 1987.

SCHERER, MATTHEW U., *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, in «Harvard Journal of Law & Technology», vol. 29, n. 2, 2015, http://dx.doi.org/10.2139/ssrn.2609777, accessed on 20.08.2024.

SEARLE, JOHN, *Minds, brains, and programs*, in «Behavioral and Brain Sciences», vol. 3, no. 3, September 1980, https://doi.org/10.1017/S0140525X00005756, pp. 417-424.

SCANTAMBURLO, TERESA, *Apprendimento automatico e decisione* umana, in *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, edited by Fabio Fossa, Viola Schiaffonati, Guglielmo Tamburrini, Carocci Editore, Roma, 2021, pp. 12-31.

SHAMS, RIFAT ARA, ZOWGHI, DIDAR, BANO, MUNEERA, *AI and the quest for diversity and inclusion: a systematic literature review*, in «AI and Ethics», 2023, https://doi.org/10.1007/s43681-023-00362-w, accessed on 20.06.2024.

SHNEIDERMAN, BEN, *Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy*, in «IUI '21 Companion: Companion Proceedings of the 26th International Conference on Intelligent User Interfaces», 2021, https://doi.org/10.1145/3397482.3453994, accessed on 30.08.2024.

SOLOMON, BRIAN, *The Hottest Startups Of 2015*, in «Forbes», 17 Dec. 2015, https://www.forbes.com/pictures/eimh45ehmdj/hottest-startups/, accessed on 14.02.2024.

SPARCK JONES, KAREN, *Natural Language Processing: A Historical Review*, in *Current Issues in Computational Linguistics: In Honour of Don Walker*, NICOLETTA CALZOLARI, MARTHA PALMER AND ANTONIO ZAMPOLLI (edited by), Springer, Dordrecht, 1994.

STIM, RICH, *Fair Use*, in *Stanford Copyright and Fair Use Center*, https://fairuse.stanford.edu/overview/fair-use/, accessed on 20.06.2024.

TADDEO, MARIAROSARIA, FLORIDI, LUCIANO, *How AI can be a force for good*, in «Science», vol. 361, n. 6404, 2018, https://doi.org/10.1126/science.aat5991, accessed on 13.04.2024.

TRONNIER, FRÉDÉRIC, ET AL., *A Systematic Literature Review on Gender Bias in AI. Towards Inclusiveness in Machine Learning*, paper presented at the Pacific Asia Conference on Information Systems (PACIS), 2024, https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1002&context=pacis2024, accessed on 23.08.2024.

TURING, ALAN MATHISON, *Computing Machinery and Intelligence*, in «Mind», volume LIX, issue 236, 1 Oct. 1950, pp. 433-460.

UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris, UNESCO, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en, accessed on 15.04.2024.

VACCARI, CRISTIAN, CHADWICK, ANDREW, *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*, in «Social Media + Society», vol. 6, 2020, https://doi.org/10.1177/2056305120903408, accessed on 15.06.2024.

VINCENT, JAMES, *Microsoft invests $1 billion in OpenAI to pursue holy grail of artificial intelligence*, in «The Verge», 22 Jul. 2019, https://www.theverge.com/2019/7/22/20703578/microsoft-openai-investment-partnership-1-billion-azure-artificial-general-intelligence-agi, accessed on 02.09.2024.

—, *OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws*, in «The Verge», 30 Jul. 2020, https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential, accessed on 02.09.2024.

VOUTIRITSAS, THEA, *10 Vivid Haikus to Leave you Breathless*, in *Read Poetry*, https://www.readpoetry.com/10-vivid-haikus-to-leave-you-breathless/, accessed on 21.08.2024.

WEIZENBAUM, JOSEPH, *ELIZA - a computer program for the study of natural language communication between man and machine*, in «Communications of the ACM», volume 9, number 1, pp. 36-45.

WITTGENSTEIN, LUDWIG, *Tractatus logico-philosophicus*, Kegan Paul, London, 1922.

XU, WEI, GAO, ZAIFENG, *Enabling Human-Centered AI: A Methodological Perspective*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.06703, accessed on 01.09.2024.

YAMASHITA, TAKASHI, MILLAR, ROBERTO J., *Likert Scale*, in DANAN GU, MATTHEW E. DUPRE (edited by), *Encyclopedia of Gerontology and Population Aging*, Cham, Springer, 2021, http://dx.doi.org/10.1007/978-3-030-22009-9, accessed on 20.03.2024.

YOGEESWARAN, KUMAR, ET AL., *The Interactive Effects of Robot Anthropomorphism and Robot Ability on Perceived Threat and Support for Robotics Research*, in «Journal of Human-Robot Interaction» vol. 5, n. 29, http://dx.doi.org/10.5898/JHRI.5.2.Yogeeswaran, accessed on 10.04.2024.

ZHENG, NAN-NING, ET AL., *Hybrid-augmented intelligence: collaboration and cognition*, in «Frontiers of Information Technology & Electronic Engineering», vol. 18, 2017, pp..153-179, https://doi.org/10.1631/FITEE.1700053, accessed on 10.04.2024.

2024 AI Index Report, https://aiindex.stanford.edu/report/, accessed on 20.08.2024.

AI Act, in European Commission, Digital Strategy, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, accessed on 20.07.2024.

Alan Turing, in *Biography*, https://www.biography.com/scientists/alan-turing, accessed on 20.02.2024.

DALL-E, in *Wikipedia*, https://en.wikipedia.org/wiki/DALL-E, accessed on 02.09.2024.

DALL·E: Creating images from text, in OpenAI, 5 Jan. 2021, https://openai.com/index/dall-e/, accessed on 13.07.2024.

Designing Likert scales, in *TASO*, https://taso.org.uk/evidence/evaluation-guidance-resources/survey-%20design-resources/evaluation-guidance-designing-likert-scales/, accessed on 20.03.2024.

Google 2024 Environmental Report, https://sustainability.google/reports/google-2024-environmental-report/, accessed on 21.08.2024.

Google, in Wikipedia, https://en.wikipedia.org/wiki/Google, accessed on 14.03.2024

Machine Learning, in *Cambridge Dictionary*, https://dictionary.cambridge.org/dictionary/english/machine-learning, accessed on 15.03.2024.

Meta Platforms, in *Wikipedia*, https://en.wikipedia.org/wiki/Meta_Platforms, accessed on 14.03.2024.

Microsoft invests in and partners with OpenAI to support us building beneficial AGI, in *OpenAI*, https://openai.com/index/microsoft-invests-in-and-partners-with-openai/, accessed on 04.09.2024.

OpenAI LP, in *OpenAI*, https://openai.com/index/openai-lp/, accessed on 03.09.2024.

Strategia italiana per l'intelligenza artificiale 2024-2026, https://www.agid.gov.it/sites/agid/files/2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-2026.pdf, accessed on 20.08.2024.

Submission Guidelines, in *Springer Nature*, https://www.nature.com/commsbio/submit/submission-guidelines, accessed on 20.08.2024.

The Evolution of Human-Computer Interaction: A Review of the Past and Future Directions, in *Association of Human-Computer Interaction*, https://www.hci.org.uk/article/the-evolution-of-human-computer-interaction-a-review-of-the-past-and-future-directions/, accessed on 20.08.2024.

Universe, in *OpenAI*, https://openai.com/index/universe/, accessed on 06.09.2024.

Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (COM/2022/496 final).

Proposal for a Directive of the European Parliament and of the Council on liability for defective products (COM/2022/495 final).

Regulation EU 2024/1689.

REFERENCES ON LARGE LANGUAGE MODELS, PROMPT ENGINEERING AND CHATGPT

ACHIAM, JOSH, ET AL., *GPT-4 Technical Report*, in *ArXiv*, 24 May 2019, https://doi.org/10.48550/arXiv.2303.08774, accessed on 15.03.2024.

ALTMAN, SAM, *Planning for AGI and beyond*, in *OpenAI*, https://openai.com/blog/planning-for-agi-and-beyond, accessed on 19.02.2024.

—, «The GPT-3 hype is way too much. It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out», in *X*, https://x.com/sama/status/1284922296348454913, accessed on 02.09.2024.

Bastian, Matthias, *GPT-4 has more than a trillion parameters – Report*, in The Decoder, 25 Mar. 2023, https://the-decoder.com/gpt-4-has-a-trillion-parameters/, accessed on 20.04.2024.

BENDER, EMILY MENON, KOLLER, ALEXANDER, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, in «Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics», Association for Computational Linguistics, 2020, pp. 5185-5198.

—, ET AL., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in «FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency», 2021, https://doi.org/10.1145/3442188.3445922, pp. 610-623.

BHATT, SHAILY , *Extrinsic Evaluation of Cultural Competence in Large Language Models*, in *ArXiv*, 19 Jun. 2024, https://doi.org/10.48550/arXiv.2406.11565, accessed on 21.08.2024.

BOLUKBASI, TOLGA, ET AL., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, in *ArXiv*, 21 Jul. 2016, https://doi.org/10.48550/arXiv.1607.06520, accessed on 29.03.2024.

BOMMASANI, RISHI, ET AL., *On the Opportunities and Risks of Foundation Models*, in *ArXiv*, 22 Jul. 2022, https://doi.org/10.48550/arXiv.2108.07258, accessed on 17.03.2024.

—, ET AL., *Do Foundation Model Providers Comply with the Draft EU AI Act?,* in *Center for Research on Foundation Models*, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html, accessed on 13.06.2024.

BOTERO ARCILA, BEATRIZ, *Is it a platform? Is it a search engine? It's ChatGPT! The European liability regime for large language models*, in «Journal of Free Speech Law», vol. 3, 2023, https://ssrn.com/abstract=4539452, accessed on 20.08.2024.

BOZKURT, ARAS, SHARMA, RAMESH C., *Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World*, in «Asian Journal of Distance Education», vol. 18, n. 2, 2023, https://doi.org/10.5281/zenodo.8174941, accessed on 20.05.2024.

BROCKMAN, GREG, *OpenAI Gym Beta*, in *OpenAI*, https://openai.com/research/openai-gym-beta, accessed on 16.02.2024

—, SUTSKEVER, ILYA, *Introducing OpenAI*, in *OpenAI*, https://openai.com/blog/introducing-openai, accessed on 14.02.2024.

BROWN, TOM B., ET AL., *Language Models are Few-Shot Learner*s, in *ArXiv*, 22 Jul. 2020, https://doi.org/10.48550/arXiv.2005.14165, accessed on 08.03.2024.

BRUNO, ALESSANDRO, ET AL., *Insights into Classifying and Mitigating LLMs' Hallucinations*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.08117, accessed on 20.08.2024.

CALISAN, AYLIN, BRYSON, JOANNA J., NARAYANAN, ARVID, *Semantics derived automatically from language corpora necessarily contain human biase*s, in *ArXiv*, 25 May 2017, https://doi.org/10.48550/arXiv.1608.07187, accessed on 13.04.2024.

CARLINI, NICHOLAS, ET AL., *Extracting Training Data from Large Language Models*, in «Proceedings of the 30th USENIX Security Symposium, 2021», https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting, accessed on 30.04.2024.

CARROLL, ALEXANDER J., BORYCZ, JOSHUA, *Integrating large language models and generative artificial intelligence tools into information literacy instruction*, in «The Journal of Academic Librarianship», vol. 50, 2024, https://doi.org/10.1016/j.acalib.2024.102899, accessed on 20.08.2024.

CHEN, BANGHAO, ET AL., *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*, in *ArXiv*, 18 Jun. 2024, https://doi.org/10.48550/arXiv.2310.14735, accessed on 22.07.2024.

CHEN, MARK, ET AL., *Evaluating Large Language Models Trained on Code*, in *ArXiv*, 14 Jul. 2021, https://doi.org/10.48550/arXiv.2107.03374, accessed on 03.09.2024.

CHEN, CANYU, SHU, Kai, *Combating Misinformation in the Age of LLMs: Opportunities and Challenges*, in *ArXiv*, 09 Nov. 2023, https://doi.org/10.48550/arXiv.2311.05656, accessed on 20.08.2024.

CHEN, ZHIYU ZOEY, et al., *A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law*, in *ArXiv*, 02 May 2024, https://doi.org/10.48550/arXiv.2405.01769, accessed on 21.08.2024.

D'AGOSTINO, ANDREA, *Cosa è ChatGPT – Approfondimento Tecnico e Consigli all'Utilizzo*, in *Diario Di Un Analista*, 14 Dec. 2023, https://www.diariodiunanalista.it/posts/cosa-e-chatgpt/, accessed on 13.03.2024.

DEVLIN, JACOB, ET AL., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in *ArXiv*, 24 May 2019, https://doi.org/10.48550/arXiv.1810.04805, accessed on 15.03.2024.

DUBEY, ABHIMANYU, ET AL., *The Llama 3 Herd of Models*, in *ArXiv*, 31 Jul. 2024, https://doi.org/10.48550/arXiv.2407.21783, accessed on 19.08.2024.

EKIN, SABIT, *Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices*, in *Authorea TechRxiv*, 04 May 2023, https://www.techrxiv.org/doi/full/10.36227/techrxiv.22683919.v1, accessed on 21.07.2024

ELOUNDOU, TYNA , ET AL., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, in *ArXiv*, 21 Aug. 2023, https://doi.org/10.48550/arXiv.2303.10130, accessed on 20.08.2024.

ENGELBART, DOUGLAS CARL, *Augmenting Human Intellect: a Conceptual Framework*, Menlo Park, Stanford Research Institute, 1962.

FANG, XIAO, ET AL., *Bias of AI-Generated Content: An Examination of News Produced by Large Language Models*, in *ArXiv*, 03 Apr. 2024, https://doi.org/10.48550/arXiv.2309.09825, accessed on 20.05.2024.

FINNE-ANSLEY, JAMES, ET AL., *My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises*, in «ACE '23: Proceedings of the 25th Australasian Computing Education Conference», https://doi.org/10.1145/3576123.3576134, accessed on 03.09.2024.

FOY, KYLIE, *AI models are devouring energy. Tools to reduce consumption are here, if data centers will adopt*, in *MIT Lincoln Laboratory*, 22 Sep. 2023, https://www.ll.mit.edu/news/ai-models-are-devouring-energy-tools-reduce-consumption-are-here-if-data-centers-will-adopt, accessed on 29.04.2024.

FRANCESCHELLI, GIORGIO, MUSOLESI, MIRCO, *On the Creativity of Large Language Models*, in *ArXiv*, 09 Jul. 2023, https://arxiv.org/abs/2304.00008, accessed on 20.08.2024.

FRIED, INA, *OpenAI says ChatGPT usage has doubled since last year*, in «Axios», 29 Aug. 2024, https://www.axios.com/2024/08/29/openai-chatgpt-200-million-weekly-active-users, accessed on 02.09.2024.

GARCÍA-MÉNDEZ, SILVIA, DE ARRIBA-PÉREZ, FRANCISCO, *Large Language Models and Healthcare Alliance: Potential and Challenges of Two Representative Use Cases*, in «Annals of Biomedical Engineering», vol. 52, 2024, pp. 1928-1931.

HUANG, BAIXIANG, CHEN, CANYU, SHU, KAI, *Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges*, in *ArXiv*, 16 Aug. 2024, https://doi.org/10.48550/arXiv.2408.08946, accessed on 20.08.2024.

HUBER, STEFAN E., et al., *Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning*, in «Educational Psychology Review», vol. 35, n. 25, 2024.

GOLDSTEIN, JOSH A., ET AL., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, in *ArXiv*, 10 Jan. 2023, https://doi.org/10.48550/arXiv.2301.04246, accessed on 15.06.2024.

GORDON, CINDY, *ChatGPT And Generative AI Innovations Are Creating Sustainability Havoc*, in «Forbes», 17 Mar. 2024, https://www.forbes.com/sites/cindygordon/2024/03/12/chatgpt-and-generative-ai-innovations-are-creating-sustainability-havoc/, accessed on 29.04.2024.

HULBERT, DAVE, *Using Tree-of-Thought Prompting to boost ChatGPT's reasoning*, in «GitHub repository», 2023, https://doi.org/10.5281/zenodo.10323452, accessed on 20.04.2024.

JONES, NICOLA, *Bigger AI chatbots more inclined to spew nonsense — and people don't always realize*, in «Nature», 25 Sep. 2024, https://www.nature.com/articles/d41586-024-03137-3, accessed on 29.09.2024.

KARAMOLEGKOU, ANTONIA, ET AL., *Copyright Violations and Large Language Models*, in «Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing», Singapore, Association for Computational Linguistics, 2023, pp. 7403-7412.

KAPLAN, JARED, ET AL., *Scaling Laws for Neural Language Models*, in *ArXiv*, 23 Jan. 2020, https://doi.org/10.48550/arXiv.2001.08361, accessed on 15.03.2024.

KE, LUOMA, ET AL., *Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review*, in *ArXiv*, 16 Mar. 2024, https://doi.org/10.48550/arXiv.2401.01519, accessed on 02.09.2024.

KELLY, SAMANTHA MURPHY, *This AI chatbot is dominating social media with its frighteningly good essays*, in «CNN Business», https://edition.cnn.com/2022/12/05/tech/chatgpt-trnd/index.html, accessed on 04.07.2024.

KIM, JINHEE, ET AL., *Exploring students' perspectives on Generative AI-assisted academic writing*, in «Education and Information Technologies», 2024, https://doi.org/10.1007/s10639-024-12878-7, accessed on 20.08.2024.

KIM, SUNNIE S. Y., ET AL., *"I'm Not Sure, But…": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust*, in *ArXiv*, 15 May 2024, https://doi.org/10.48550/arXiv.2405.00623, accessed on 29.08.2024.

KOJIMA, TAKESHI, ET AL., *Large Language Models are Zero-Shot Reasoners*, in ArXiv, 29 Jan. 2023, https://doi.org/10.48550/arXiv.2205.11916, accessed on 19.03.2024.

GG, AOBO, ET AL., *Better Zero-Shot Reasoning with Role-Play Prompting*, in *ArXiv*, 15 Mar. 2024, https://doi.org/10.48550/arXiv.2308.07702, accessed on 24.05.2024.

LAWRENCE, NEIL, *OpenAI won't benefit humanity without data-sharing*, in «The Guardian», 14 Dec. 2015, https://www.theguardian.com/media-network/2015/dec/14/openai-benefit-humanity-data-sharing-elon-musk-peter-thiel, accessed on 15.02.2024.

LEE, UNGGI, ET AL., *Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education*, in «Education and Information Technologies», vol. 29, 2024, https://doi.org/10.1007/s10639-023-12249-8, accessed on 1.05.2024.

LI, CHENG, ET AL., *CultureLLM: Incorporating Cultural Differences into Large Language Models*, in *ArXiv*, 09 Feb. 2024, https://doi.org/10.48550/arXiv.2402.10946, accessed on 21.08.2024.

LI, CHUNYUAN , ET AL., *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*, in *ArXiv*, 01 Jun. 2023, https://doi.org/10.48550/arXiv.2306.00890, accessed on 02.09.2024.

LI, JIAYIANG, LI, JIALE, SU, YUNSHENG, *A Map of Exploring Human Interaction Patterns with LLM: Insights into Collaboration and Creativity*, in HELMUT DEGEN, STAVROULA NTOA (edited by), *Artificial Intelligence in HCI*, vol. 14735, Cham, Springer, 2024, https://doi.org/10.1007/978-3-031-60615-1_5, accessed on 02.09.2024.

LIN, ZICHAO, ET AL., *Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models*, in «Artificial Intelligence Review», vol. 57, n. 243, 2024, https://doi.org/10.1007/s10462-024-10896-y, accessed on 20.08.2024.

LIYANAGE, UDARA PIYASENA , RANAWEERA, NIMNAKA DILSHAN, *Ethical Considerations and Potential Risks in the Deployment of Large Language Models in Diverse Societal Contexts*, in «Journal of Computational Social Dynamics«, vol. 8, n. 11, 2023, https://vectoral.org/index.php/JCSD/article/view/49, accessed on 21.08.2024.

LIU, PENGFEI, ET AL., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, in *ArXiv*, 28 Jul. 2021, https://doi.org/10.48550/arXiv.2107.13586, accessed on 15.07.2024.

LUCCHI, NICOLA, *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, in «European Journal of Risk Regulation», 2023, https://doi.org/10.1017/err.2023.59, accessed on 23-07.2024.

MANJOO, FARHAD, *How Do You Know a Human Wrote This?*, in «The New York Times», 29 Jul. 2020, https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html, accessed on 04.09.2024.

METZ, CADE, *OpenAI Lets Mom-and-Pop Shops Customize ChatGPT*, in «The New York Times», 06 Nov. 2023, https://www.nytimes.com/2023/11/06/technology/openai-custom-chatgpt.html, accessed on 09.09.2024.

—, *The New Chatbots Could Change the World. Can You Trust Them?*, in «The New York Times», 11 Dec. 2022, https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html, accessed on 04.05.2024.

—, *ChatGPT Can Now Generate Images, Too*, in «The New York Times», 06 Nov. 2023, https://www.nytimes.com/2023/09/20/technology/chatgpt-dalle3-images-openai.html, accessed on 09.09.2024.

MIN, BONAN, ET AL., *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*, in *ArXiv*, 1 Nov. 2021, https://doi.org/10.48550/arXiv.2111.01243, accessed on 15.03.2024.

MINAEE, SHERVIN, ET AL., *Large Language Models: A Survey*, in *ArXiv*, 20 Feb. 2024, https://doi.org/10.48550/arXiv.2402.06196, accessed on 20.07.2024.

MITCHELL, MELANIE, *Can Large Language Models Reason?*, in *AI: A Guide for Thinking Humans*, 5 Dec. 2022, https://aiguide.substack.com/p/can-large-language-models-reason, accessed on 20.02.2024.

NAVEED, HUMZA, ET AL., *A Comprehensive Overview of Large Language Models*, in *ArXiv*, 09 Apr. 2024, https://doi.org/10.48550/arXiv.2307.06435, accessed on 14.07.2024.

NAYAK, SHIVA PRASAD, ET AL., *GDPR Compliant ChatGPT Playground*, in «2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)», 2024, pp. 1-6, http://dx.doi.org/10.1109/ICETCS61022.2024.10543557, accessed on 07.08.2024.

NING, LIN, ET AL., *User-LLM: Efficient LLM Contextualization with User Embeddings*, in *ArXiv*, 21 Feb. 2024, https://doi.org/10.48550/arXiv.2402.13598, accessed on 29.07.2024.

OPENAI, *Terms of use*, https://openai.com/policies/row-terms-of-use/, accessed on 20.08.2024.

OUYANG, LONG, ET AL., T*raining language models to follow instructions with human feedback*, in *ArXiv*, 4 Mar. 2022, https://doi.org/10.48550/arXiv.2203.02155, accessed on 20.03.2024.

PAPAGEORGIOU, ELEFTHERIA, et al., *A Survey on the Use of Large Language Models (LLMs) in Fake News*, in «Future Internet», vol. 15, n. 8, 2023, https://doi.org/10.3390/fi16080298, accessed on 20.08.2024

PARK, YE-JEAN, et al., *Assessing the research landscape and clinical utility of large language models: a scoping review*, in «BMC Medical Informatics and Decision Making», vol. 24, 2024, https://doi.org/10.1186/s12911-024-02459-6, accessed on 20.08.2024.

PATEL, DYLAN, WONG, GERALD, *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*, https://www.semianalysis.com/p/gpt-4-architecture-infrastructure, accessed on 15.02.2024.

PATTERSON, DAVID, ET AL., *Carbon Emissions and Large Neural Network Training*, in *ArXiv*, 23 Apr. 2021, https://doi.org/10.48550/arXiv.2104.10350, accessed on 20.07.2024.

PEARCE, HAMMOND, ET AL., *Examining Zero-Shot Vulnerability Repair with Large Language Models*, in *ArXiv*, 15 Aug. 2022, https://doi.org/10.48550/arXiv.2112.02125, accessed on 03.09.2024.

PERRIGO, BILLY, *AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits*, in «Time», 05 Dec. 2022, https://time.com/6238781/chatbot-chatgpt-ai-interview/, accessed on 20.03.2024.

—, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, in «Time», 18 Jan. 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/, accessed on 07.04.2024.

POPE, AUDREY, *NYT v. OpenAI: The Times's About-Fac*e, in «Harvard Law Review», 10 Apr. 2024, https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/, accessed on 20.06.2024.

POPLI, NIK, *GPT-4 Has Been Out for 1 Day. These New Projects Show Just How Much More Powerful It Is*, in «Time», 15 Mar. 2023, https://time.com/6263475/gpt4-ai-projects/, accessed on 10.09.2024

PORTER, JON, *ChatGPT continues to be one of the fastest-growing services ever*, in «The Verge», 6 Nov. 2023, https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference, accessed on 14.02.2024.

RADFORD, ALEC, ET AL., *Improving Language Understanding by Generative Pre-Training*, in OpenAI, 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, accessed on 04.05.2024.

—, *Language Models are Unsupervised Multitask Learners*, in *OpenAI*, 2019, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, accessed on 03.09.2024.

REYNOLDS, LARIA, MCDONELL, KYLE, *Multiversal views of language models*, in *ArXiv*, 12 Feb. 2021, https://doi.org/10.48550/arXiv.2102.06391, accessed on 27.02.2024.

ROBERTS, JESSE, *How Powerful are Decoder-Only Transformer Neural Models?*, in *ArXiv*, 02 Feb. 2024, https://doi.org/10.48550/arXiv.2305.17026, accessed on 02.09.2024.

ROBERTSON, DONALD, *FSF-funded call for white papers on philosophical and legal questions around Copilot: Submit before Monday, August 23, 2021*, in *Free Software Foundation*, 28 Jul. 2021, https://www.fsf.org/blogs/licensing/fsf-funded-call-for-white-papers-on-philosophical-and-legal-questions-around-copilot, accessed on 03.09.2024.

ROBISON, KYLIE, *OpenAI's new model is better at reasoning and, occasionally, deceiving*, in «The Verge», 17 Sep. 2024, https://www.theverge.com/2024/9/17/24243884/openai-o1-model-research-safety-alignment, accessed on 24.09.2024.

RONG, XIN, *word2vec Parameter Learning Explained*, in *ArXiv*, 11 Nov. 2014, https://doi.org/10.48550/arXiv.1411.2738, accessed on 27.02.2024.

ROOSE, KEVIN, *The Brilliance and Weirdness of ChatGPT*, in «The New York Times», 1 Mar. 2023, https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html, accessed on 15.02.2024.

—, *The New ChatGPT Can 'See' and 'Talk.' Here's What It's Like*, in «The New York Times», 27 Sep. 2023, https://www.nytimes.com/2023/09/27/technology/new-chatgpt-can-see-hear.html, accessed on 09.09.2024.

ROSENBAUM, ERIC, *AI is getting very popular among students and teachers, very quickly*, in «CNBC», 11 Jun. 2024, https://www.cnbc.com/2024/06/11/ai-is-getting-very-popular-among-students-and-teachers-very-quickly.html, accessed on 28.09.2024.

SCHARTH, MARCEL, *The ChatGPT chatbot is blowing people away with its writing skills. An expert explains why it's so impressive*, in «The Conversation», 06 Dec. 2022, https://theconversation.com/the-chatgpt-chatbot-is-blowing-people-away-with-its-writing-skills-an-expert-explains-why-its-so-impressive-195908, accessed on 12.05.2024.

SCHWARTZ, OSCAR, *Could 'fake text' be the next global political threat?*, in «The Guardian», 04 Jul. 2019, https://www.theguardian.com/technology/2019/jul/04/ai-fake-text-gpt-2-concerns-false-information, accessed on 04.09.2024.

SCHWINN, LEO, ET AL., *Adversarial Attacks and Defenses in Large Language Models: Old and New Threats*, in *ArXiv*, 20 Dec. 2023, https://doi.org/10.48550/arXiv.2310.19737, accessed on 18.04.2024.

SHAHRIAR, SAKIB, ET AL., *Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency*, in *ArXiv*, 19 Jun. 2024, https://doi.org/10.48550/arXiv.2407.09519, accessed on 09.09.2024.

SHANAHAN, MURRAY, *Talking About Large Language Models*, in *ArXiv*, 7 Dec. 2022, https://doi.org/10.48550/arXiv.2212.03551, accessed on 27.02.2024.

SKJUVE, MARITA, FØLSTAD, ASBJØRN, BRANDTZAEG, PETTER BAE, *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*, in «CUI '23: ACM conference on Conversational User Interfaces», 2023, https://doi.org/10.1145/3571884.3597144, accessed on 10.04.2024.

TANG, KUNSHENG, et al., *GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models*, in *ArXiv*, 22 Aug. 2024, https://doi.org/10.48550/arXiv.2408.12494, accessed on 23.08.2024.

TEMSAH, REEM, ET AL., *Healthcare's New Horizon With ChatGPT's Voice and Vision Capabilities: A Leap Beyond Text*, in «Cureus», vol. 15, 2023, https://doi.org/10.7759/Fcureus.47469, accessed on 02.09.2024.

TIULKANOV, ALEKSANDR, «A simple algorithm to decide whether to use ChatGPT, based on my recent article», in *X*, https://x.com/shadbush/status/1616007675145240576, accessed on 02.04.2024.

TOKAYEV, KASSYM-JOMART, *Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns*, in «International Journal of Social Analytics», vol. 8, n. 3, 2023, https://norislab.com/index.php/ijsa/article/view/42, accessed on 20.08.2024.

TSENG, YU-MIN, ET AL., *Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization*, in *ArXiv*, 26 Jun. 2024, https://doi.org/10.48550/arXiv.2406.01171, accessed on 19.08.2024.

TUBELLA, ANDREA ALER, MORA-CANTALLOPS, MARÇAL, NIEVES, JUAN CARLOS, *How to teach responsible AI in Higher Education: challenges and opportunities*, in «Ethics and Information Technology», vol. 26, n. 3, 2023, https://doi.org/10.1007/s10676-023-09733-7, accessed on 08.08.2024.

UNESCO, *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. in *UNESCO*, 7 Mar. 2024, https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes, accessed on 20.04.2024.

VASWANI, ASHISH, et al., *Attention is all you need*, paper presented at the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017.

VILLALOBOS, PAOLO, ET AL., *Will we run out of data? Limits of LLM scaling based on human-generated data*, in *ArXiv*, 26 Oct. 2022, https://doi.org/10.48550/arXiv.2211.04325, accessed on 15.03.2024.

VYKOPAL, IVAN, ET AL., *Disinformation Capabilities of Large Language Models*, 23 Feb. 2024, in *ArXiv*, https://doi.org/10.48550/arXiv.2311.08838, accessed on 15.06.2024.

WANG, BOSHI, ET AL., *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters*, in «Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics», vol. 1, pp. 2717-2739, https://aclanthology.org/2023.acl-long.153, accessed on 04.07.2024.

WANG, LI, ET AL., *Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs*, in «Npj Digital Medicine», vol. 7, n. 41, 2024, https://doi.org/10.1038/s41746-024-01029-4, accessed on 20.07.2024.

WANG, TIANYU, ZHOU, NIANJUN, CHENG, ZHIXIONG, *Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation*, in *ArXiv*, 07 Jul. 2024, https://doi.org/10.48550/arXiv.2407.05437, accessed on 20.07.2024.

WANG, XUEZHI, ET AL., *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, 07 Mar. 2023, https://doi.org/10.48550/arXiv.2203.11171, accessed on 09.05.2024.

WANG, YIXU, ET AL., *Fake Alignment: Are LLMs Really Aligned Well?*, in *ArXiv*, 10 Nov. 2023, https://doi.org/10.48550/arXiv.2311.05915, accessed on 20.02.2024.

WARZEL, CHARLIE, *The Most Important Job Skill of This Century*, in «The Atlantic», 8 Feb. 2023, https://www.theatlantic.com/technology/archive/2023/02/openai-text-models-google-search-engine-bard-chatbot-chatgpt-prompt-writing/672991/, accessed on 20.07.2024.

WEI, JASON, et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, in *ArXiv*, 10 Jan. 2023, https://doi.org/10.48550/arXiv.2201.11903, accessed on 29.04.2024.

—, *Finetuned language models are zero-shot learners*, in *ArXiv*, 08 Feb. 2022, https://doi.org/10.48550/arXiv.2109.01652, accessed on 23.04.2024.

WEI, ALEXANDER, HAGHTALAB, NIKA, STEINHARDT, JACOB, *Jailbroken: How Does LLM Safety Training Fail?*, in *ArXiv*, 05 Jul. 2023, https://doi.org/10.48550/arXiv.2307.02483, accessed on 08.03.2024.

WEIL, ELIZABETH, *You Are Not a Parrot And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this*, in «Journal of Applied Learning & Teaching», 1 Mar. 2023, https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html, accessed on 19.02.2024.

WHITE, JULES, et al., *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, in *ArXiv*, 21 Feb. 2023, https://doi.org/10.48550/arXiv.2302.11382, accessed on 13.04.2024.

WU, TIANYU, ET AL., *A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development*, in «IEEE/CAA Journal of Automatica Sinica», vol. 10, no. 5, 2023, pp. 1122-1136.

YANG, DIYI, WU, SHERRY TONGSHUANG, HEARST, MARTI A., *Human-AI Interaction in the Age of LLM*s, in «Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies», vol. 5, pp. 34-38.

YANG, JINGFENG, ET AL., *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*, in «ACM Transactions on Knowledge Discovery from Data», vol. 18, n. 160, pp. 1-32, https://doi.org/10.1145/3649506, accessed on 28.09.2024.

YAO, SHUNYU, ET AL., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, in *ArXiv*, 03 Dec. 2023, https://doi.org/10.48550/arXiv.2305.10601, accessed on 05.04.2024.

YE, JUNJIE, ET AL., *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models*, in *ArXiv*, 23 Dec. 2023, https://doi.org/10.48550/arXiv.2303.10420, accessed on 10.09.2024.

YIN, SHUKANG, ET AL., *A Survey on Multimodal Large Language Models*, in *ArXiv*, 1 Apr. 2024, https://doi.org/10.48550/arXiv.2306.13549, accessed on 15.03.2024.

ZAREMBA, WOJCIECH, BROCKMAN, GREG, *OpenAI Codex*, in *OpenAI*, 10 Aug. 2021, https://openai.com/index/openai-codex/, accessed on 03.09.2024.

ZHANG, CHIYUAN, ET AL., *Counterfactual Memorization in Neural Language Models*, in *ArXiv*, 13 Oct. 2023, https://doi.org/10.48550/arXiv.2112.12938, accessed on 20.03.2024.

ZHANG, DUZHEN, ET AL., *MM-LLMs: Recent Advances in MultiModal Large Language Models*, in *ArXiv*, 28 May 2024, https://doi.org/10.48550/arXiv.2401.13601, accessed on 02.09.2024.

ZHANG, LIANG, CHEN, ZHELUN, *Opportunities and Challenges of Applying Large Language Models in Building Energy Efficiency and Decarbonization Studies: An Exploratory Overview*, in *ArXiv*, 18 Dec. 2023, https://doi.org/10.48550/arXiv.2312.11701, accessed on 09.05.2024.

ZHAO, HAIYAN, ET AL., *Explainability for Large Language Models: A Survey*, in *ArXiv*, 28 Nov. 2023, https://doi.org/10.48550/arXiv.2309.01029, accessed on 21.04.2024.

ZHAO, WAYNE XIN, ET AL., *A Survey of Large Language Models*, in *ArXiv*, 24 Nov. 2023, https://doi.org/10.48550/arXiv.2303.18223, accessed on 17.03.2024.

ZHOU, YONGCHAO, ET AL., *Large Language Models Are Human-Level Prompt Engineers*, in *ArXiv*, 10 Mar. 2023, https://doi.org/10.48550/arXiv.2211.01910, accessed on 03.08.2024.

ZHU, WENHAO , ET AL., *Extrapolating Large Language Models to Non-English by Aligning Languages*, in *ArXiv*, 09 Oct. 2023, https://doi.org/10.48550/arXiv.2308.04948, accessed on 21.08.2024.

ZOU, ANDY, ET AL., U*niversal and Transferable Adversarial Attacks on Aligned Language Models*, in *ArXiv*, 20 Dec. 2023, https://doi.org/10.48550/arXiv.2307.15043, accessed on 18.04.2024.

AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice, in *European Commission Digital Strategy*, https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice, accessed on 28.09.2024.

ChatGPT can now see, hear, and speak, in *OpenAI*, 25 Sep. 2023, https://openai.com/index/chatgpt-can-now-see-hear-and-speak/, accessed on 09.09.2024.

ChatGPT: OpenAI riapre la piattaforma in Italia garantendo più trasparenza e più diritti a utenti e non utenti europei, in *Garante per la Protezione dei Dati Personali*, https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490, accessed on 20.05.2024.

ChatGPT plugins, in *OpenAI*, 23 Mar. 2023, https://openai.com/index/chatgpt-plugins/, accessed on 09.09.2024.

Gpt-2, in *GitHub*, https://github.com/openai/gpt-2, accessed on 02.09.2024.

GPT-2: 1.5B release, in *OpenAI*, 5 Nov. 2019, https://openai.com/index/gpt-2-1-5b-release/, accessed on 20.08.2024.

Few-Shot Prompting, in *Prompt Engineering Guide*, https://www.promptingguide.ai/techniques/fewshot, accessed on 13.04.2024.

Guidance to civil servants on use of generative AI, in *United Kingdom Government*, https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai/guidance-to-civil-servants-on-use-of-generative-ai, accessed on 07.08.2024.

Hello GPT-4o, in *OpenAI*, 13 May 2024, https://openai.com/index/hello-gpt-4o/, accessed on 04.09.2024.

How Gemini for Google Cloud uses your data, in *Google Cloud*, https://cloud.google.com/gemini/docs/discover/data-governance, accessed on 29.07.2024.

How your data is used to improve model performance, in *OpenAI*, https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance, accessed on 29.07.2024.

Introducing ChatGPT, in *OpenAI*, 30 Nov. 2022, https://openai.com/index/chatgpt/, accessed on 11.05.2024.

Introducing OpenAI o1-preview, in *OpenAI*, 12 Sep. 2024, https://openai.com/index/introducing-openai-o1-preview/, accessed on 24.09.2024.

Introduction to Large Language Models, in *Google Cloud Skills Boost*, https://www.cloudskillsboost.google/course_templates/539, accessed on 14.03.2024.

Large language model, in Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/large-language-model, accessed on 14.03.2024.

Learning to Reason with LLMs, in *OpenAI*, 12 Sep. 2024, https://openai.com/index/learning-to-reason-with-llms/, accessed on 24.09.2024.

OpenAI, in *Wikipedia*, https://en.wikipedia.org/wiki/OpenAI, accessed on 14.03.2024.

What are large language models (LLMs)?, in *IBM*, https://www.ibm.com/topics/large-language-models, accessed on 17.03.2024.

REFERENCES ON ARTIFICIAL INTELLIGENCE IN EDUCATION AND ARTIFICIAL INTELLIGENCE LITERACY

BAIDOO-ANU, DAVID, OWUSU ANSAH, LETICIA, *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*, in «Journal of AI», vol. 7, 2023, https://doi.org/10.61969/jai.1337500, accessed on 20.05.2024.

BEARMAN, MARGARET, AJJAWI, ROLA, *Learning to work with the black box: Pedagogy for a world with artificial intelligence*, in «British Journal of Educational Technology», 21 Nov. 2023, https://theconversation.com/why-student-experiments-with-generative-ai-matter-for-our-collective-learning-210844, accessed on 03.03.2024.

BORENSTEIN, JASON, HOWARD, AYANNA, *Emerging challenges in AI and the need for AI ethics education*, in «AI and Ethics», vol.1, 2021, pp. 61-65, https://doi.org/10.1007/s43681-020-00002-7, accessed on 15..05.2024

ČERNÝ, MICHAL, *University Students' Conceptualisation of AI Literacy: Theory and Empirical Evidence*, in «Social Sciences», vol. 13, n. 3, 2024, https://doi.org/10.3390/socsci13030129, accessed on 29.08.2024.

CETINDAMAR, DILEK, ET AL., *Explicating AI Literacy of Employees at Digital Workplaces*, in «IEEE Transactions of Engineering Management», http://dx.doi.org/10.1109/TEM.2021.3138503, accessed on 05.09.2024.

OPARA, EMMANUEL CHINONSO, ADALIKWU MFON-ETTE, THERESA, TOLORUNLEKE, CAROLINE ADUKE, *ChatGPT for Teaching, Learning and Research: Prospects and Challenges*, in «Global Academic Journal of Humanities and Social Sciences», vol. 5, pp. 33-40, https://ssrn.com/abstract=4375470, accessed on 10.07.2024.

EUROPEAN COMMISSION, *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*, Bruxelles, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2766/153756, accessed on 13.08.2024.

—, *AI report – By the European Digital Education Hub's Squad on artificial intelligence in education*, Bruxelles, Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2797/828281, accessed on 13.08.2024.

FEDERAL MINISTRY OF EDUCATION AND RESEARCH (BMBF), *Richtlinie zur Bund-Länder-Initiative zur Förderung der Künstlichen Intelligenz in der Hochschulbildung*, 2021, retrieved from https://www.bmbf.de/bmbf/shareddocs/bekanntmachungen/de/2021/02/3409_bekanntmachung.html , accessed on 02.09.2024.

FERSTER, BILL, *Teaching Machines. Learning from the Intersection of Education and Technology*, Maryland, John Hopkins University Press, 2014, p. 1.

HAUGBAKEN, HADVAN, HAGELIA, MARIANNE, *A New AI Literacy For The Algorithmic Age: Prompt Engineering Or Educational Promptization?*, in «2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)», 2024, http://dx.doi.org/10.1109/ICAPAI61893.2024.10541229, accessed on 15.06.2024.

KEYHANI, MOHAMMAD, ET AL., *Why student experiments with Generative AI matter for our collective learning*, in «The Conversation», vol. 54, no. 5, 2023, pp. 1160-1173.

KNOTH, NILS, ET AL., *AI literacy and its implications for prompt engineering strategies*, in «Computers and Education: Artificial Intelligence», vol. 6, 2024, https://doi.org/10.1016/j.caeai.2024.100225, accessed on 20.08.2024.

KONG, SIU-CHEUNG, CHEUNG, WILLIAM MAN-YIN, ZHANG, GUO, *Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds*, in «Computers and Education: Artificial Intelligence», vol. 2, 2021, https://doi.org/10.1016/j.caeai.2021.100026, accessed on 05.06.2024.

LAUPICHLER, MATTHIAS CARL, ET AL., *Artificial intelligence literacy in higher and adult education: A scoping literature review*, in «Computers and Education: Artificial Intelligence», vol. 3, 2022, https://doi.org/10.1016/j.caeai.2022.100101, accessed on 03.03.2024.

LO, LEO S., *The CLEAR path: A framework for enhancing information literacy through prompt engineering*, «The Journal of Academic Librarianship», vol. 49, 2023, https://doi.org/10.1016/j.acalib.2023.102720, accessed on 13.04.2024.

LONG, DURI, MAGERKO, BRIAN, *What is AI Literacy? Competencies and Design Considerations*, in «CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems», pp. 1-16, https://doi.org/10.1145/3313831.3376727, accessed on 20.04.2024.

MARKAUSKAITE, LINA, ET AL., *Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?*, in «Computers and Education: Artificial Intelligence», vol. 3, 2022, https://doi.org/10.1016/j.caeai.2022.100056, accessed on 13.04.2024.

NG, DAVY TSZ KIT, ET AL., *Conceptualizing AI literacy: An exploratory review*, in «Computers and Education: Artificial Intelligence», vol. 2, 2021, https://doi.org/10.1016/j.caeai.2021.100041, accessed on 02.05.2024.

PARK WOOLF, BEVERLY, ET AL., *AI Grand Challenges for Education*, in «AI Magazine», vol. 34, 2013, https://doi.org/10.1609/aimag.v34i4.2490, accessed on 13.04.2024.

PRENSKY, MARC, *Digital Natives, Digital Immigrants*, in «On the Horizon», vol. 9, 2001, http://dx.doi.org/10.1108/10748120110424816, accessed on 13.04.2024.

RUDOLPH, JÜRGEN, ET AL., *ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?*, in «Journal of Applied Learning & Teaching», vol. 6, no. 1, 2023, https://doi.org/10.37074/jalt.2023.6.1.9, accessed on 14.02.2024.

SALVAGNO, MICHELE, TACCONE, FABIO SILVIO, GERLI, ALBERTO GIOVANNI, *Correction to: Can artificial intelligence help for scientific writing?*, in «Critical Care», vol. 27, n. 99, 2023, https://doi.org/10.1186/s13054-023-04390-0, accessed on 20.08.2024.

SULLIVAN, MIRIAM, ET AL., *Improving students' generative AI literacy: A single workshop can improve confidence and understanding*, in «Journal of Applied Learning & Teaching», vol. 7, n. 2, 2024, http://journals.sfu.ca/jalt/index.php/jalt/index, accessed on 10.09.2024.

TAN, CHEE WEI, *Large Language Model-Driven Classroom Flipping: Empowering Student-Centric Peer Questioning with Flipped Interaction*, in *ArXiv*, 14 Nov. 2023, https://doi.org/10.48550/arXiv.2311.14708, accessed on 15.06.2024.

THEOPHILOU, EMILY, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, paper presented at the 22nd International Conference of the Italian Association for Artificial Intelligence (AIXIA), Rome, Italy, 2023.

UNESCO, MIAO, FENGHCUN, HOLMES, WAYNE, *Guidance for generative AI in education and research*, Paris, UNESCO, 2023, https://doi.org/10.54675/EWZM9535, accessed on 15.03.2024.

UNIMORE, PRESS OFFICE , *EDUNEXT. Al via il progetto per l'innovazione della formazione digitale a livello nazionale*, University of Modena and Reggio Emilia, 2024, https://www.magazine.unimore.it/site/home/notizie/articolo820069931.html, accessed on 09.09.2024.

WALTER, YOSHIJA, *Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education*, in «International Journal of Educational Technology in Higher Education», vol. 21, n. 15, 2024, https://doi.org/10.1186/s41239-024-00448-3, accessed on 21.07.2024.

WANG, TIANYU, ZHOU, NIANJUN, CHENG, ZHIXIONG, *Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation*, in *ArXiv*, 07 Jul. 2024, https://doi.org/10.48550/arXiv.2407.05437, accessed on 20.07.2024.

WINGARD, JASON, *ChatGPT: A Threat To Higher Education?*, in «Forbes», 10 Jan. 2023, https://www.forbes.com/sites/jasonwingard/2023/01/10/chatgpt-a-threat-to-higher-education/, accessed on 02.03.2024.

WOO, DAVID JAMES, et al., *Effects of a Prompt Engineering Intervention on Undergraduate Students' AI Self-Efficacy, AI Knowledge, and Prompt Engineering Ability: A Mixed Methods Study*, in *ArXiv*, 30 Jul. 2024, https://doi.org/10.48550/arXiv.2408.07302, accessed on 20.08.2024.

ZAMFIRESCU-PEREIRA, J. D., ET AL., *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*, in «Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems», 2023, https://doi.org/10.1145/3544548.3581388, accessed on 15.03.2024.

ZHAI, CHUNPENG, WIBOWO, SANTOSO, LI, LILY D., *The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review*, in «Smart Learning Environments», vol. 11, n. 28, 2024, https://doi.org/10.1186/s40561-024-00316-7, accessed on 26.09.2024.

Abilità informatiche: nuovo corso sull'IA per lauree umanistiche e lingue, in *Ca' Foscari University of Venice*, 03 Sep.2024, https://unive.it/pag/14024/?tx_news_pi1%5Bnews%5D=15770&cHash=4eebcb25d37fc994bb7619b7308c4c55, accessed on 05.09.2024.

Introducing ChatGPT Edu, in *OpenAI*, 20 May 2024, https://openai.com/index/introducing-chatgpt-edu/, accessed on 02.09.2024.

APPENDIX

LIST OF FIGURES

PROJECT CHARTER

Name	An empirical study on how artificial intelligence literacy and prompt engineering influence the use of LLMs and GAI within higher education
Referents	Valentina Rossi Professor Teresa Scantamburlo
Questions	What is the effect of literacy on students in terms of formulating effective prompts? What is the effect of literacy on students in terms of their opinion and knowledge of GAI and LLMs? How does literacy contribute to students' ethical awareness of the use of GAI?
Scope of work	Design, implementation and evaluation of a literacy workshop in generative artificial intelligence (GAI) and prompt engineering for non-technical students, in order to assess the impact on interactions and perceptions
Expected outcomes	Improved skills in prompt engineering, Increased awareness of the importance of this knowledge in the academic and professional worlds, Increased awareness of the ethical issues and limitations of LLMs and GAI, Possible framework for the effective integration of these concepts
Performance metrics	Analysis of the questionnaire (Likert scale) Analysis of the prompts and the outputs

Phases	Phase 1: pre-questionnaire and preliminary exercise Phase 2: Literacy on GAI, LLMs, and prompt engineering Phase 3: post-questionnaire and final exercise
Participants	9 students from humanities faculties
Literacy topics	Definition of ‘intelligent’ and artificial intelligence Brief history of GAI, and LLMs Architecture of LLMs/ChatGPT (in general) Limitation and biases Ethical, social and legal challenges Successful case studies and examples of safe and productive utilisation Prompt engineering techniques and prompt patterns

QUESTIONNAIRE

Pre-questionnaire (30 questions, 3 exercises)

1. Gender, age, nationality (3 questions)
2. 1 practical exercise
3. Previous experience with AI, GAI and ChatGPT (9 questions)
4. Initial perceived level of threat (7 questions)[466]
5. Interaction quality evaluation: ChatGPT capabilities, Effort perceived to achieve desired behaviour (9 questions)[467]
6. Ethical considerations: ethical awareness, future utility (9 questions)

Post-questionnaire (45 questions, 3 exercises)

1. 1 practical exercise
2. Final perceived level of threat (7 questions)[468]
3. Emotions resulting from interaction (8 questions)[469]
4. Interaction quality evaluation: ChatGPT capabilities, Effort perceived to achieve desired behaviour (9 questions)[470]
5. ChatGPT as tool: pragmatic dimension, hedonic dimension, human likeness, social

[466] The questions have been translated from the previous study E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

[467] Some questions have been adapted from two previous studies: E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit., and M. SKJUVE, A. FØLSTAD, P. BAE BRANDTZAEG, *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*, cit.

[468] The questions have been translated from the previous study E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

[469] The questions have been translated from the previous study E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

[470] Some questions have been adapted from the previous study E. THEOPHILOU, ET AL., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit.

presence (14 questions)[471]

6. Perceived interaction improvement (3 questions)

7. Ethical considerations: ethical awareness, future utility (9 questions)

Gender, age, nationality (*pre-questionnaire*)	
What genre do you identify with?	Male, Female, Non-binary
What is your age?	18-20, 20-22, 22-24, 24+
What is your nationality?	Italian, European, Other

Previous experience with AI, GAI and ChatGPT (*pre-questionnaire*)	
How familiar are you with the concept of artificial intelligence (AI)?	No knowledge, Limited knowledge, Medium knowledge, Good knowledge, Very good knowledge
How familiar are you with the concept of generative artificial intelligence (GAI)?	
Have you used ChatGPT-3.5 or ChatGPT-4o (the free versions) before?	Never, Once or twice, A few times, Very often, I use it almost everyday
Have you used ChatGPT-3.5 or ChatGPT-4o (the free versions) for academic purposes (e.g. to find ideas, facilitate exam preparation, facilitate essay writing)?	
Have you used ChatGPT-4 (the paid version) before?	
Have you used ChatGPT-4 (the paid version) for academic purposes (e.g. to find ideas, facilitate exam preparation, facilitate essay writing)?	
How familiar are you with the concept and techniques of prompt engineering?	No knowledge, Limited knowledge, Medium knowledge, Good knowledge, Very good knowledge

[471] Some questions have been adapted from two previous studies: E. Theophilou, et al., *Learning to Prompt in the Classroom to Understand AI Limits: A pilot study*, cit., and M. Skjuve, A. Følstad, P. Bae Brandtzaeg, *The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users*, cit.

Have you used other generative artificial intelligence (GAI) tools, such as text-to-image, text-to-speech, etc.?	Never, Once or twice, A few times, Very often, I use it almost everyday
If so, which ones?	[...]
Perceived level of threat (*pre-questionnaire*, *post- questionnaire*)	
Artificial intelligence applications are beginning to blur the boundaries between human and machine	From 1 (Strongly disagree) to 5 (Strongly agree)
The increased use of artificial intelligence in our lives is causing humans to lose their jobs	
Artificial intelligence implementations can effectively replace workers from their jobs	
In the long run, artificial intelligence poses a direct threat to human welfare and safety	
The realism of artificial intelligence is disturbing because it makes it almost indistinguishable from human beings	
Recent advances in artificial intelligence are challenging the very essence of what it means to be human	
Technological advances in artificial intelligence are threatening the uniqueness of humans	
Emotions resulting from interaction (*post-questionnaire*)	
Anger	From 1 (Strongly disagree) to 5 (Strongly agree)
Fear	
Disgust	
Anxiety	
Sadness	
Desire	

Serenity	
Joy	
Interaction quality evaluation (*pre-questionnaire, post-questionnaire*) **ChatGPT capabilities**	
I am satisfied with ChatGPT's comprehension and response capabilities	From 1 (Strongly disagree) to 5 (Strongly agree)
ChatGPT demonstrates to be intelligent	
ChatGPT repeats the same mistakes over and over again, without adapting to my questions	
ChatGPT demonstrates an understanding of complex concepts	
ChatGPT demonstrates human-like reasoning and comprehension	
Effort perceived to achieve desired behaviour	
I found it easy to communicate my intentions to ChatGPT	From 1 (Strongly disagree) to 5 (Strongly agree)
Obtaining desired responses from ChatGPT required an acceptable level of effort on my part	
I had to repeat my questions or requests multiple times to get satisfactory responses from ChatGPT	
Interacting with ChatGPT required more effort than I initially expected	
ChatGPT as a tool (*post-questionnaire*) **Pragmatic dimension**	
ChatGPT enables efficiency and enhances quality of work	From 1 (Strongly disagree) to 5 (Strongly agree)

ChatGPT provides relevant and useful output	
ChatGPT does not want to answer the request due to policies or because it only has limited information in its database	
ChatGPT presents misinformation or biased views	
Hedonic dimension	
ChatGPT is exceeding expectations, impressive, or superior compared to existing solutions	From 1 (Strongly disagree) to 5 (Strongly agree)
ChatGPT can support creative activities (such as essay writing, brainstorming or dialectical exchanges)	
Interactions with ChatGPT are entertaining	
Interactions with ChatGPT stimulate intellectual curiosity and engagement	
Human likeness	
ChatGPT is humanlike or intelligent	From 1 (Strongly disagree) to 5 (Strongly agree)
ChatGPT's responses seem to come from a real person	
ChatGPT exhibits human-like reasoning and comprehension	
Social presence	
I feel like I was engaged in an active dialogue with ChatGPT	From 1 (Strongly disagree) to 5 (Strongly agree)
Interactions with ChatGPT feel like a conversation between equals, where we naturally answer each other's questions	
I feel as if ChatGPT and I are involved in a common task when interacting	

Perceived interaction improvement (*post-questionnaire*)	
Since participating in the literacy course, I have noticed an improvement in my ability to interact effectively with ChatGPT	From 1 (Strongly disagree) to 5 (Strongly agree)
The literacy course has helped me better understand how to engage with ChatGPT for more meaningful interactions	
After completing the literacy course, I feel easier to get useful responses from ChatGPT	
Ethical considerations (*pre-questionnaire*, *post-questionnaire*) **Ethical awareness**	
I am conscious of the potential biases embedded in large language models (LLMs)	From 1 (Strongly disagree) to 5 (Strongly agree)
I believe it is important for users to understand the ethical implications of using large language models (LLMs), like ChatGPT	
I feel responsible for ensuring that my interactions with generative artificial intelligence (GAI) tools adhere to ethical guidelines and principles	
I believe in the importance of regularly reviewing and updating ethical guidelines for the use of large language models (LLMs) and generative artificial intelligence (GAI)	
I recognise the potential for large language models (LLMs) like ChatGPT to perpetuate harmful stereotypes or misinformation	
Future utility	
Prompt engineering techniques will become indispensable for navigating advanced AI systems in the future	From 1 (Strongly disagree) to 5 (Strongly agree)

Mastery of prompt engineering will empower humans to harness the full potential of AI systems for problem-solving and innovation

The integration of prompt engineering skills into academic and professional practices will lead to more efficient and impactful outcomes

AI literacy will foster critical thinking skills necessary for evaluating and ethically engaging with AI technologies in various contexts

*The questions written in italics were specifically created for this study.

PROMPT ENGINEERING EXERCISE

Deliverable

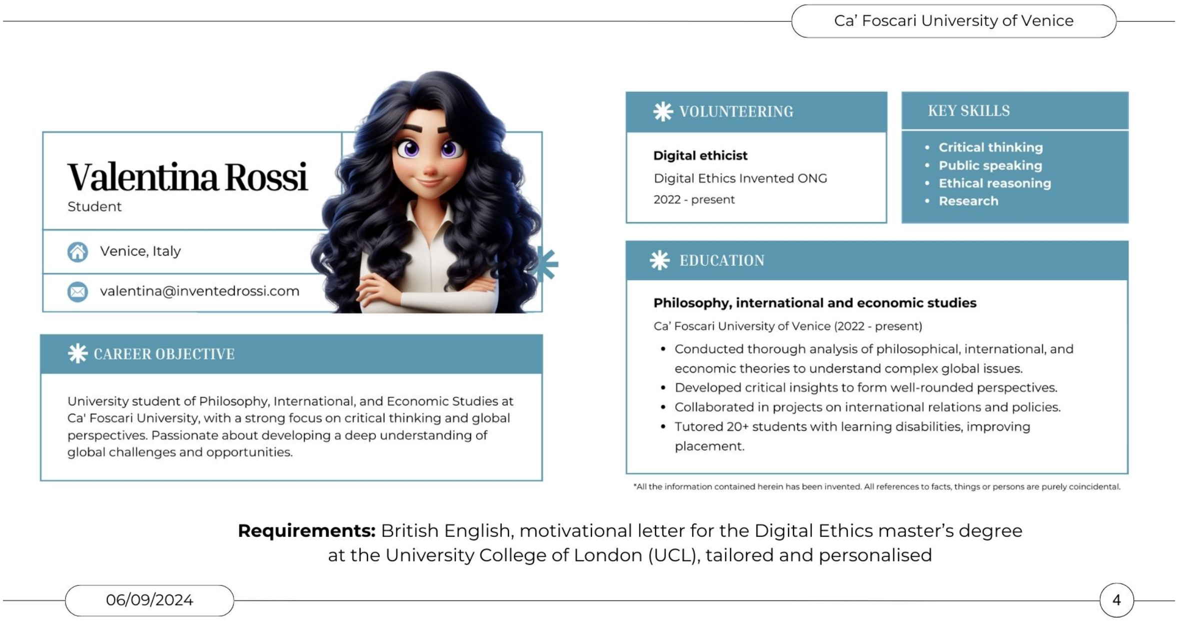
Here below you are provided with a CV of a fictional person containing totally invented information.

You will have to interact with ChatGPT in order to obtain a motivational letter for possible admission to a Master's degree course at the University College of London (UCL) in Digital Ethics.

The letter has to be in British English. You can interact with up to 5 inputs with the chatbot (in a single conversation), however the end result must be a motivational letter in format and text structure.

When you have finished, you will be asked to share the conversation (you will find instructions under the CV).

*Remember that this is not an assessment, but an exercise to see how we can improve!



Ca' Foscari University of Venice

Valentina Rossi
Student

Venice, Italy

valentina@inventedrossi.com

CAREER OBJECTIVE

University student of Philosophy, International, and Economic Studies at Ca' Foscari University, with a strong focus on critical thinking and global perspectives. Passionate about developing a deep understanding of global challenges and opportunities.

VOLUNTEERING

Digital ethicist
Digital Ethics Invented ONG
2022 - present

KEY SKILLS

- **Critical thinking**
- **Public speaking**
- **Ethical reasoning**
- **Research**

EDUCATION

Philosophy, international and economic studies
Ca' Foscari University of Venice (2022 - present)

- Conducted thorough analysis of philosophical, international, and economic theories to understand complex global issues.
- Developed critical insights to form well-rounded perspectives.
- Collaborated in projects on international relations and policies.
- Tutored 20+ students with learning disabilities, improving placement.

*All the information contained herein has been invented. All references to facts, things or persons are purely coincidental.

Requirements: British English, motivational letter for the Digital Ethics master's degree at the University College of London (UCL), tailored and personalised

06/09/2024

4

FORMULAS, CHAPTER I.1.

Recurrent neural networks (RNNs)

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1})$$

h_t is the hidden state at time step t

σ is a nonlinear activation function, often the sigmoid function or the hyperbolic tangent ($tanh$) function

W_{xh} is the weight matrix that connects the input x_t to the hidden state

x_t is the input vector at time step t

W_{hh} is the weight matrix connecting the hidden state from the previous time step h_{t-1} to the current hidden state

h_{t-1} is the hidden state from the previous time step

Multi-head attention

$$\text{head}_i = attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$multihead(Q, K, V) = \text{concat}(head_1, head_2, \ldots, head_h)W^0$$

$W_i^Q W_i^K W_i^V$ are specific weight matrices for the head i, used to project Q, K, and Vin different dimensional spaces for each attention head

$concat$ after calculating the attention of each head separately, the results (the various heads) are concatenated together

W^0 is a final weight matrix that projects the concatenated output into a desired output space.

Self-attention mechanism

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Q (query) is the vector representing the information to be retrieved, as for the case of transformers, each word in a textual sequence can have its own query vector

K (key) is the vector used to compare against the query y: each token in the textual sequence has also a key vector. Specifically, it is used to determine how relevant each word in the sequence is to the query.

V (value) is the vector containing the actual information associated with each word, i.e. the value used to inform the model's decision on the output

d_k is the dimensionality of the keys and queries

$softmax$ is the function converting values to probabilities, emphasising higher values (indicating stronger links between queries and keys)

$\left(\frac{QK^T}{\sqrt{d_k}}\right)$ this scalar product is used to calculate the similarity between each query and key vector, normalizing the result to avoid extreme values: it measures how relevant each element in the sequence is to the current query, preparing the data for application of the SoftMax function.