UNIVERSITA' DEL SALENTO

Corso di Laurea in Management Digitale

Tesi di Laurea in Machine Learning in Management

Integrating Generative AI and LLMs in Business: Harnessing Python for Data-Driven Decision Making and Predictive Analytics

RELATORE:

Ch.mo Prof. Fabrizio Antonio, Durante

STUDENTE:

Matteo Scarciglia

Matricola n° 20063198

Anno Accademico 2022 - 2023

Summary

Riassunto

Il qui presente lavoro intende esplorare le fondamenta teoriche, le applicazioni pratiche e più specificamente l'impatto dell'Intelligenza Artificiale Generativa all'interno dei contesti aziendali. Quanto proposto è una dimostrazione molto pratica di un Sistema Informativo AI-powered, il quale, grazie all'implementazione tecnica di un Modello di Linguaggio di Grandi Dimensioni (LLM), è ora predisposto per un'interazione aumentata con i suoi utilizzatori, indipendentemente dal livello organizzativo cui essi appartengono o dalle specifiche funzioni aziendali. Questo sistema permetterebbe un nuovo approccio all'accesso all'informazione, non solo più efficace ma anche e soprattutto facilitato dalla possibilità di poter interrogare dati strutturati o destrutturati utilizzando soltanto il linguaggio naturale, come nel parlato comune; ogni utente coinvolto nel business riuscirebbe potenzialmente ad ottenere benefici per mezzo di questa maggiore "alfabetizzazione informativa", grazie al quale utenti con un più basso livello di hard-skills, ad esempio, possono scalare il prodotto delle proprie attività lavorative verso un livello superiore, oppure grazie al maggior senso di appartenenza ed empowerment generato dal sapere meglio su cui si lavora. Essendo questo sistema essenzialmente illimitato nelle possibilità di sviluppo, di customizzazione e di scalabilità, la versatilità sarebbe impattante praticamente in ogni macro e micro-funzione organizzativa.

Ciò che aggiunge ulteriore carattere innovativo al presente lavoro è la dimostrazione di come, mediante l'utilizzo del linguaggio di programmazione Python e di alcune specifiche librerie, utenti di business che non sono verticalizzati nel ramo dell'informatica possano sfruttare una conoscenza generata da parte di una vastissima community di esperti, di altri sistemi di low-code, o ancora di AI, per accedere facilmente a risorse di elevatissimo valore e addirittura gratuite, e creare sistemi più o meno complessi e personalizzati, dotati di una Graphical User Interface (GUI). Il potere di creare sistemi di questo tipo da parte di persone direttamente coinvolte nel business, permette non solo di eliminare qualsiasi barriera dovuta al differente approccio al business interposta tra l'utilizzatore finale

del sistema informativo e il team di informatici coinvolti nel suo sviluppo, ma permetterebbe anche, soprattutto laddove la richiesta di un sistema di Business Intelligence di questo tipo è caratterizzata da bassa complessità, di abbattere sensibilmente, grazie ad un minor approvvigionamento di competenze specifiche.

Introduction

The aim of the present work is to explore the theoretical foundations and practical applications of Generative Artificial Intelligence and the underlying Large Language Models (LLM) within business contexts. What has been argued is that through the use of new forms of Information Systems based on LLM (such as advanced machine learning models), capable of integrating data and associated information as never before, access to information can be greatly facilitated for all actors involved in the business, resulting in innumerable advantages in organisational operations. What adds further innovative character to the present work is the demonstration of how, through the use of the Python programming language and a few specific libraries, business users who are not vertically specialised in the branch of information technology can exploit the knowledge generated by a vast community of experts in an extremely versatile manner, easily accessing extremely valuable and even free resources, to create more or less complex, customised systems, and equipped with a Graphical User Interface (GUI). The power to create systems of this type by people directly involved in the business allows not only to eliminate any macro or micro barrier due to the different approach to the business in terms of skills interposed between the end user of the information system and the management team computer scientists involved in its development, but it would also allow, where the request for a Business Intelligence system of this type is characterized by low complexity, to reduce costs and operate more economically, thanks to the cost reduction resulting from a lower provision of specific skills.

The work begins with a discussion on the theoretical foundations of Generative Artificial Intelligence and LLMs in which an overview of the underlying machine learning algorithms is provided. Then it shows how these models are capable of generating a wide range of data, such as images, audio, and text. This theoretical approach to the subject will be turned into a practical application through the creation of an AI-based assistant accessible via a web app programmed in Python, in which the interface and back-end operation, libraries and frameworks will be shown for application development, API integration into the system, and action capabilities, including benefits, limitations, and prompt-effectiveness.

In the following, the impact on company operations is analysed through an exploration of the possibilities for developing an "enhanced" information system and the analysis of a new AI-powered business model taxonomy, followed by an analysis of the costs of implementation and development based on the benefits obtained. Alongside this, the main functional and technological features of this system will be described, which will serve both to facilitate its study from a point of view that is no longer just technical, but also relating to focal aspects such as data management, or even from a legal point of view.

The issue of privacy, considering its sensitivity, is the subject of further in-depth analysis, which proposes the different regulatory approaches adopted by Nations, and how heterogeneous cultural perimeters deal in different ways with the issue of the protection of natural and legal persons, with reference to the personal data concerning them.

The conclusions instead leave further food for thought with respect to the potential evolutions of this sub-branch of science and the essentiality of understanding the economic repercussions and challenges posed by the ongoing revolution, placing emphasis on the importance of anticipating and navigating the transformations induced by Generative AI in the business management sector.

Chapter 1) Theoretical Foundations of a New General-Purpose Technology – What Generative AI is and how it works

The ability of Generative AI of affecting an entire economy, drastically altering societies through its impact is the characteristic which most makes Generative AI a new General-Purpose Technology, just like electricity or internet itself. Generative AI could be understood as a complex system of machine learning algorithms capable of learning large amounts of information to train a computer, which, when queried by a user in natural language, is able to generate text content, images, audio, lines of code or other. This is a simply yet powerful way to define what Artificial Intelligence is and what can it do for people, enabling new ways of understanding the working world, education, research, or simply everyday life. But before diving deeper in what Generative AI can do for humans a look on its roots is needed, in order to understand what it was supposed to do as first, and how it has evolved compared to the target. In the 1940s and 1950s several scientists from various fields of science began to question whether or not it was possible to recreate the human brain through the use of electronics. This concept of "electronic brain" can be attributed to the seed that some philosophers, physics, mathematicians, engineers, and neurologists planted in an attempt to describe the thinking process of the human brain as mechanical manipulation of symbols. In that period, there were several attempts to formalize these ideas and lay the groundwork to concretely advance with new research in various areas of the sciences. The intuition that a new form of brain capable of being not unlike a human being arose from the fusion of theories in the subjects of neurology, computer and information science, philosophy and mathematics. It all, in fact, started by considering how the human brain was an electrical network of neurons that communicated with each other via a dichotomous signal, that is, on or off. Clearly, this agreed perfectly with Alan Turing's computational theory (Wikipedia, n.d.), which showed how any problem can be solved by an algorithm; this perfectly embraced Claude Shannon's information theory (Wikipedia, n.d.), according to which any form of computation can be described digitally. Indeed, it was intuited how these theories were compatible and unified by much scientific evidence, and it was said how these, combined, could enable the construction of an "electronic brain". Formally, the real foundations of research in the field of artificial intelligence in its primordial form were laid in the second half of the 1950s by Dartmouth College in the famous "Dartmouth workshop" held by Marvin Minsky, John McCarthy, Claude Shannon and Nathan Rochester, who wanted to revive the conceptual definition of Artificial Intelligence reiterating how “every aspect of learning or any other feature of intelligence can be so precisely describer that a machine can be made to simulate it”. In the wake of this new wave of thought and enthusiasm that was being created among experts in the field, various scientists introduced different approaches to this phenomenon. Among the most influential approaches were Reasoning as Search, Neural Networks or natural language, Natural Language Processing (NLP) and reinforcement learning. All these practices had, in their own way, important implications in the field of the search to the most appropriate algorithm, and all had their successes and failures. In the midst of these machine learning techniques, however, one theory proved particularly suitable for use for this purpose, namely, Artificial Neural Networks (ANN)[1]. This machine learning techniques consist of interconnected groups of nodes, akin to biological neurons, which collectively process information. The fundamental unit of these networks is the artificial neuron, or node, which receives input, processes it, and transmits output to other neurons. Each node is interconnected with other nodes before and after and has an associated weight value that it will use to weight (activate or not activate) the output of the previous node and that will be the input of the next node.

[1] More information available on (International Business Machines Corporation, n.d.)

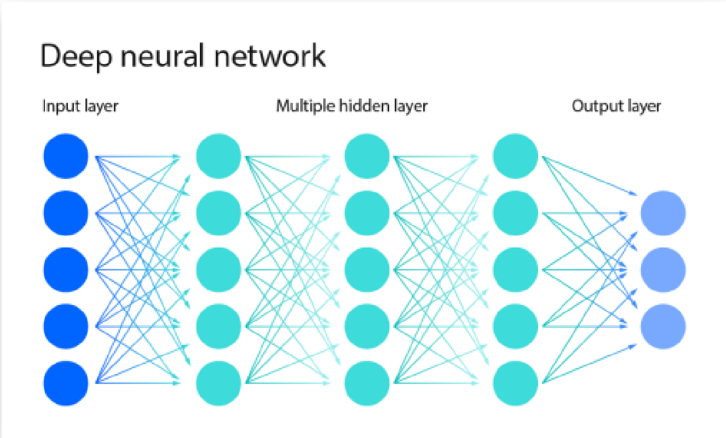


Figure 1 - Composition of an Artificial Neural Network (Source: www.ibm.com - Article on neural networks).

The concept of neural networks was first introduced in the 1940s by Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician. Their work laid the foundation for ANNs by demonstrating how neural circuits in the brain could compute logical functions. Since the 1950s, several possible types of Neural Networks (Ferrario C., 2022)[2] have been studied; among them are:

i) **Perceptron**[3] (Keim R., 2019): introduced in the late 1950s by psychologist Frank Rosenblatt; it is presented as a type of binary classifier that maps its inputs x to an output value $f(x)$ calculated with

$$f(x) = 1(< w, x > + b)$$

[2] Ferrario C., 2022, "Reti neurali artificiali: cosa sono, come funzionano e perché vengono costruite", Geopop, https://www.geopop.it/reti-neurali-artificiali-cosa-sono-come-funzionano-e-perche-vengono-costruite/ of 20 November 2022.

[3] Keim R., 2019, "How to Train a Basic Perceptron", All About Circuits, https://www.allaboutcircuits.com/technical-articles/how-to-train-a-basic-perceptron-neural-network/ of 24 November 2019.

where w is a vector of weights with real values, the operator <.,.> is the scalar product (which calculates a weighted sum of the *inputs*), b is the *bias*, a constant term that does not depent on any *input* value, and $1(y)$ is the *output* function; that is set to 1 when the condition in its argument is satisfied, and 0, otherwise;

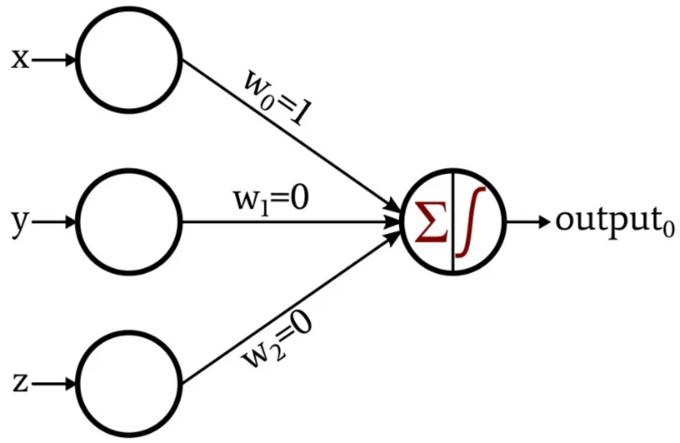


Figure 2 - Graphical representation of a perceptron (Source: https://www.allaboutcircuits.com/technical-articles/how-to-train-a-basic-perceptron-neural-network/).

ii) **Feed Forward Networks**: this type of network is characterised by the passage of information travelling in a single direction. This type of networks is characterised by the presence of only one input layer and one output layer (single-layer networks), or by the presence of intermediate layers of hidden nodes if they are multi-layer (Tosato M., 2013)[4]. Each layer has connections with the previous and the next, and within these the signal propagates forward without transverse connections. Such a network can be seen as a set of perceptrons:

---

[4] Tosato M., 2013, "Addestramento reti neurali feed-forward multi-layered tramite Error Backpropagation", Paperblog, https://it.paperblog.com/addestramento-reti-neurali-feed-forward-multi-layered-tramite-error-backpropagation-1988250 of 30 April 2013.

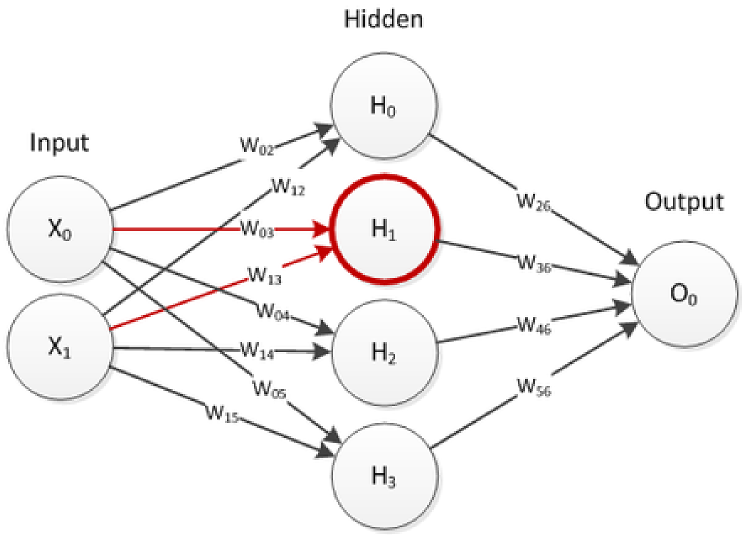


Figure 3 - Graphical representetion of a Neural feed-forward multi-layered Network (Source: https://it.paperblog.com/addestramento-reti-neurali-feed-forward-multi-layered-tramite-error-backpropagation-1988250/).

iii) **Recurrent Neural Networks (RNN)**: this type of neural network, unlike the previous one, is characterised by the possibility of activating perceptrons in a bi-directional manner, implying that the output of a node can be the input of the same node.

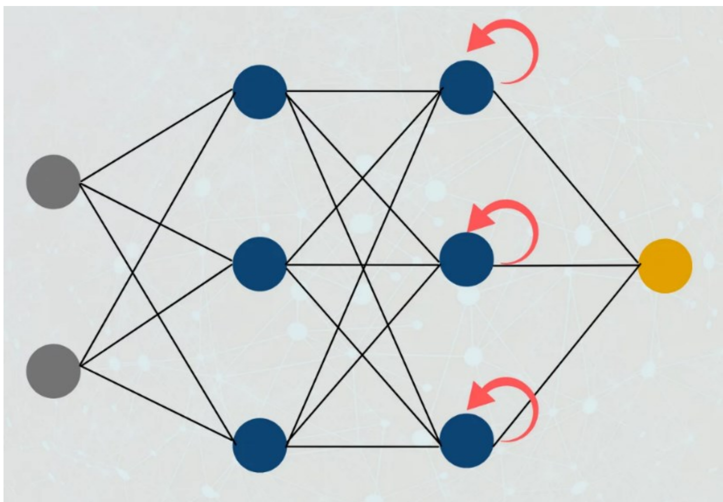


Figure 4 - Graphical representation of a Recurrent Neural Network (Source: https://databasecamp.de/en/ml/recurrent-neural-network).

These algorithms are defined (DatabaseCamp, 2021)[5], together with the previous one, as Deep Learning and are characterised by their ability to use training data to learn, as if they had a memory. The fields of use range from language translation to Natural Language Processing (NLP), to speech recognition and image detection applications.

i) **Convolutional Neural Networks (CNN or ConvNet)**: this type of neural network (Fan X., 2022)[6] is a special kind of multilayer feed forward network, however, consisting of more than five layers. All layers that are not input and output, i.e. the so-called 'convulsive layers' are hidden and perform a mapping function of the features received as input. This resulting map will be used as input information for the next layer to arrive at a detailed output.

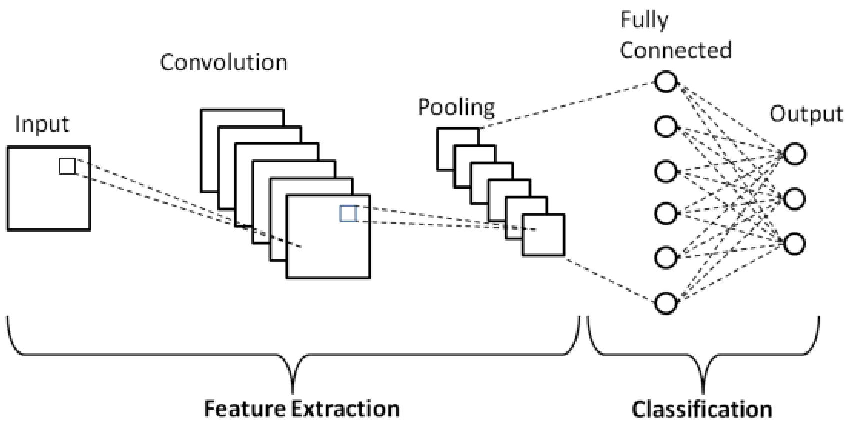


Figure 5 - Graphical representation of a Convultional Neural Network (Source: https://medium.com/sfu-cspmp/an-introduction-to-convolutional-neural-network-cnn-207cdb53db97).

---

[5] 2021, "What are Recurrent Neural Networks?", Data Base Camp, https://databasecamp.de/en/ml/recurrent-neural-network of 14 December 2021.
[6] Fan X., Truong P., 2022, "An Introduction to Convolutional Neural Network (CNN)", Medium, https://medium.com/sfu-cspmp/an-introduction-to-convolutional-neural-network-cnn-207cdb53db97 of 11 February 2022.

The use of this type of neural network is particularly popular in image and video recognition systems, recommendation systems and bioinformatics.

Actually, the path that led to the emergence of these algorithms, faced a major problem due to the lack of computing power required to bring these algorithms to work on a large scale. Furthermore, another major issue that prevented their development was the absence of common knowledge fixed on appropriate media. Each of these algorithms, in fact, requires an enormous amount of data on which to be trained, and without this information, their usefulness cannot be exploited.

The field of Artificial Intelligence has attracted attention in a fluctuating manner, alternating between periods of strong interest and periods of neglect. It was only after about 50 years, with the new era of IT, that a more concrete interest began to develop, this time however supported by numerous scientific proofs. This success was mainly due to the increase in the computing power of computers, which finally allowed more consistent training of these models, who began to tackle with the right weapons problems that, although simple, were nevertheless specific. One of the milestones considered most important for the development of AI is the birth of Deep Blue[7] (Feng-hsiung H., 1999) the IBM supercomputer that managed to beat chess player Garry Kasparov in 1997 in a six-game re-match, achieving two wins and three draws; this increase in computing power was surprisingly respecting Moore's Law[8] (Riccò B., 2008), which predicted that the speed and memory of computers would double every two years, thanks to the development of ever more innovative and technologically advanced transistors. Following this trend, which made the production of computers more and more powerful and less and less expensive, coupled with the development of huge new sources of data through the production of big data, has been possible to see the use of more advanced machine learning algorithms, now capable of solving increasingly complex problems intertwined

[7] Murray C., Hoane A. J. Jr., Feng-hsiung H., 2002, "Deep Blue", Artificial Intelligence.
[8] Riccò B., 2008, "Legge di Moore", Treccani, http://www.treccani.it/enciclopedia/legge-di-moore_%28Enciclopedia_della_Scienza_e_della_Tecnica%29/.

within the real economy. These advancements, in fact, have enabled Artificial Intelligence systems to a transformative progression, heralding a new epoch of digital augmentation, where AI penetrates diverse domains of human activity. Future research on the implications of this technological proliferation, specifically the economic impact, is integral to understand and navigate the potentials and challenges this digital revolution presents.

1.1) Overview of Machine Learning algorithms behind Large Language Models (LLMs)

Within the AI landscape, Generative AI fits as an alternative to other statistical tools such as unsupervised learning, supervised learning, and reinforcement learning equally complex but simpler to use. More precisely, Generative AI working model is based on supervised learning models, which are trained using different ways of programming on labelled datasets, where they learn to map input data to the correct output. This learning paradigm is foundational for understanding and predicting the structure inherent in the data, as demonstrated previously with neural networks. It plays a pivotal role in the training phase of generative models, where the AI is often first trained in a supervised manner to understand the basic structure and distribution of the data before it starts generating new instances. Thus, Generative AI uses algorithms such as neural networks to train based on output, implementing a minimization of the cost function between actual and predicted outcome. Some algorithms such as Recurrent Neural Networks, in fact, are among the algorithms capable of self-training and mimicking the concept of human memory, to understand what should be repeated over time and what should not. What turns out is that the decade of around 2010-2020 was a decade of large-scale supervised learning, and this is very important since laid the foundation for modern Generative AI. Starting around 2010, a lot of data started to be generated from a lot of applications and devices, but even if researchers fed the models more data, its performance was not getting that much better compared to the training of a small AI model. But after some time and as the research progressed, researchers started to realize through this period that a large AI model trained on a very large amount on data, was getting better and better as the model size, in terms of power and memory, increases too. Due to the capability of these new systems, this time powered by major computational resources, many applications have born, like spam detection, weather forecasting, image classification, face recognition and so on.

Today, with Generative AI all society refers to a system able to transform verbose human intention, submitted via prompt, in some useful and re-usable results, with features of originality, curiosity, creativity, as a humans would have done. This generation process is guaranteed by a technology, which allows Generative AI to generate texts: Large Language Model (LLM). This technology can understand and generate text based on the text the user provide in with a prompt. The way LLMs do this is, given an input like "I love eating" (prompt), the model complete this sentence with some options it got in during the training phase. This happens because of their nature to be a supervised learning-type technique, so that the model, has learned to label a certain input with the corresponding output, like "pasta and pizza" in this case. So if the model has read on the Internet a sentence like this, then this sentence will be turned into a lot of data points for it to try to learn to predict the next word. Thus, training a very large AI system on a lot of data, like trillion words, then the result is a Large Language Model that given a prompt is very good at generating some additional words in response to that prompt. So, this discretionary space left in the hands of the algorithm allows for answers that, just as a human would, taking into account his values and experience, range between the various data points created in training and thus generate responses contextually relevant, reflecting a wide spectrum of knowledge and perspectives gleaned from the training data.

The working model of LLMs is based on the representation of words in multidimensional vectors, which are able to represent words in such a way that those with similar contextual meanings or other relationships are approximated in vector space. Basically, a Neural Network formed by different Layers and Connections uses the Transformer to be trained (supervised) on different Datasets and, finally, to Generate and Yield and answer. The key component of a LLM is the Transformer[9], which is a deep learning model introduced in 2017 by an article named "Attention Is All You Need" (Vaswani A. et al., 2017) and who revolutionized the way of intending Natural Language Processing (NLP). The fundamental characteristics of a transformer are inherent in its architecture; the architecture of a transformer is based on the presence of multiple layers, each containing a feed-forward neural network, starting with an encoder and ending with a decoder. These feed-forward networks allow the transformer to process the input sequence in parallel and simultaneously, unlike RNNs which can only process data sequentially. This processing works through the mechanism of self-attention, whereby it can differentially weight the significance of each part of the sentence and 'pay attention' to different parts of an input sequence in different ways.

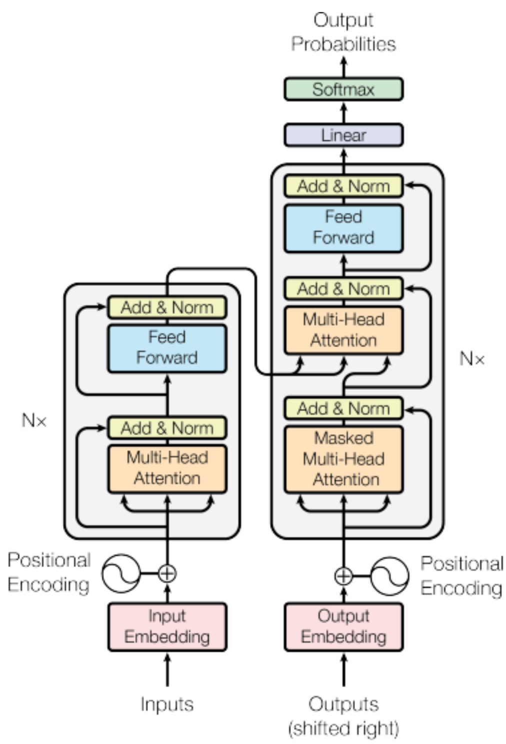


Figure 6 - The Transformer - Model architecture (Source: https://arxiv.org/pdf/1706.03762.pdf).

[9] Vaswani A. et al., 2017, "Attention Is All You Need", Cornell University, https://arxiv.org/abs/1706.03762 of 12 June 2017.

The underlying layers of large neural networks that are based on transformers are interconnected by the nodes of the network, which orientate the input of the preceding node after weighing it for a parameter, which changes during training (a large neural network can have billions of parameters). The training of a transformer takes place by giving input huge amounts of high-quality data; during training, the model then iteratively adjusts the parameters of each individual node until it correctly predicts the sequence of tokens, thus creating a data point. Some example of tasks LLMs can carry out are Writing, Reading or Answering a question. Given a training on specific argument, the LLM can be good at writing a specific text content based on the provided information, reading the information used in training phase, and answering to specific questions also. For example, it can read the customer emails of an online shop and help the manager to quickly figure out the content and set up a response. Another practical example of LLMs application is their usage in chatbots; ChatGPT, Bard and Microsoft Copilot are some examples of general-purpose chatbots, where use LLMs to act as a Generative AI system ready to provide every information user need, requested via a web interface-based or software-based application.

Remarkable as they are, however, LLMs are not infallible. As an example, being trained on information generated by humans, reflects the quality of these information, resulting in a Biased output. These biases can be racial, gender, cultural or ideologic based and, combined to a not-understanding totally the real context like humans do, LLMs yield answer based on learned pattern, without understanding the actual significance (e.g. can not be able to understand irony, sarcasm or other kind of humourist). For example, in a paper named "A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia and Youtube"[10] (Luo, Puett, & Smith, 2023), which is studied the cultural and language perception in the context of search engine and online platform

[10] Queenie L., Puett M., Smith M., 2023, "A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and Youtube", Cornell University, https://arxiv.org/abs/2303.16281 of 28 March 2023.

like Google, Wikipedia or ChatGPT, is demonstrated that these platforms influence the way users perceive complex arguments and the representation of different cultures. What is demonstrated is that, in the case of search engine already, different search results are shown to users depending on which language do they speak, affecting the information they can gather and lastly the idea they can build of an argument. In the case of LLM, ChatGPT is taken in exam and what researchers found is that being this trained mainly in English, appear to be "almost blind" to not-anglophonic perspectives; so, the document says the way ChatGPT yields results is biased by the Anglo-American perspective, reducing the complexity of a multi-faced question to this unique standard. Regarding ChatGPT, researchers found that, to reduce this biases, LLM's training data "should be a multilingual dataset in its origin, which each language having a reasonably balanced representation, and outputs answers in the user's input language through automatic translation". Moreover, an automatic translation system is proposed, to allow ChatGPT to answer in the user input language so that users can receive information relevant to its culture and language, reducing biases. Other issues related to LLMs could be dangerous content creation like fake news, environmental problems considering the energy they require to work, or unrealistic expectance for users, since users can misunderstand the capacity of the model and misuse the system.

Chapter 2) Practical Application: Building an AI Assistant through a Web-app coded using Python

Understood how many possibilities there are with a system like those above, it is also remarkable how could be easy to build a system tailored to what are the needs and expectations of the user. Not only the advancement of these technology is noticeable, but the ease of use also. In recent years, a notable shift has been observed in the landscape of software-based and web-based application development. This transformation is largely attributed to advancements in technology and the proliferation of tools, such as programming languages and consequently pre-built modules, libraries and functions, designed to simplify the app development process. These changes, mainly due to the ecosystem born across communities made up by researchers online, which gave their contributions freely publishing e diffusing a lot of open-source content and technical material, significantly lowered the barriers to entry, enabling individuals with limited programming experience to engage in the creation of applications. Firstly, the emergence of no-code and low-code platforms has been a pivotal factor in democratizing the applications' development. As identified in a report by Gartner[11] (Vincent P. et al., 2021) these platforms offer visual development environments, enabling user to construct applications using drag-and-drop interfaces and model-driven logic through a graphical user interface. Moreover, the advent of user-friendly development tools and frameworks has also contributed to this trend. In web-applications especially, technologies like React[12], developed and maintained by Meta, and Angular, developed and maintained by Google, have introduced more intuitive ways of building web applications. They, in fact, encapsulate complex functionalities into easier-to-use components, thus lowering the learning curve for novice programmers. As mentioned, lastly, the role of community-driven learning cannot be understated. Platforms such as Stack

[11] Vincent P., Wong J., Saikat R., Akash J., Guttridge K., Kimihiko I., Leow A., Natis Y., 2021, "Identify and Evaluate Your Next Low-Code Development Technologies", Gartner, Vincent P., Wong J., Saikat R., Akash J., Guttridge K., Kimihiko I., Leow A., Natis Y. of 13 April 2021.

[12] More information available React and Angular available here: https://react.dev/learn and https://angular.io/.

Overflow and GitHub have fostered collaborative environments where novice developers can learn from contribute to a vast repository of shared knowledge and code. This communal approach to learning and development has played a crucial role in enabling individuals with limited programming background to venture into applications development[13] (Juhás G. et al., 2022).

As a demonstration of what has just been argued, an example of a web-app, built by a non-expert programmer (actually, me, a management undergraduate), will be shown during this chapter, thanks to what these changes in the dissemination of know-how are developing. Precisely, the web-app is intended to demonstrate how it is possible to build, in a relatively short time and at practically no cost (except for a few euro cents), an application capable of integrating such an innovative technology as generative AI and making it usable for non-experts. The development of this application took about four months, including time for training (greatly facilitated for the reasons mentioned above, due in part to the fact that all the frameworks used, which will be described below, are open source), studying the user experience and debugging; the main objective was to train an LLM on specific data held by the user, so that the user could query the data via a simple prompt. The interface is equipped with a text space, in which the user can enter his prompt, and a space above it containing the response taken directly from the data provided by the LLM, which will also be referred to below as the 'Agent'. The application is now accessible via prompt only since it is hosted on a local virtual server not public available due to the low deployment level of this library. Later, more deploy-oriented libraries will be shown.

---

13 G. Juhás, L. Molnár, A. Juhásová, M. Ondrišová, M. Mladoniczky and T. Kováčik, 2022, "Low-code platforms and languages: the future of software development", IEEE Xplore, Low-code platforms and languages: the future of software development | IEEE Conference Publication | IEEE Xplore.

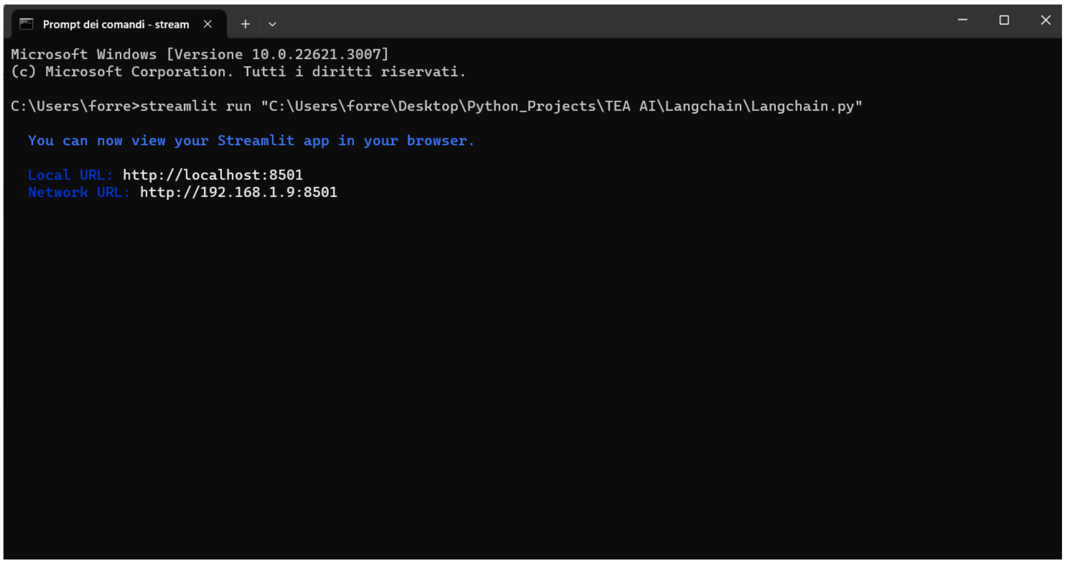


Figure 7 - Running of the application, hosted on a local private virtual server.

This system is intended to be easy to use, minimizing the difference between the user experience of a program that the average computer user is able to use and a more complex system. The GUI (Graphical User Interface) is designed to be as usable as possible:

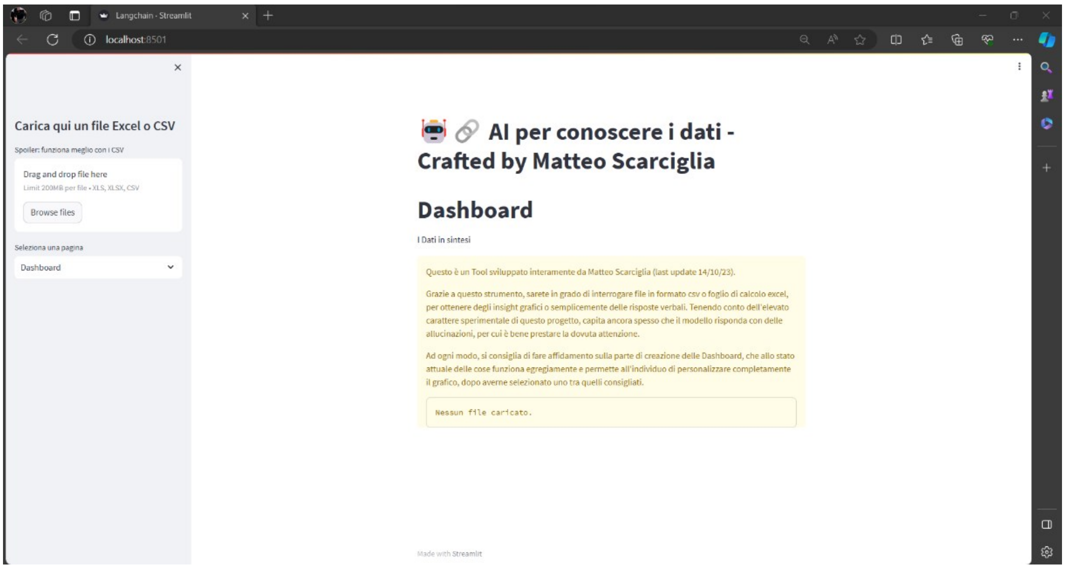


Figure 8 - Graphical User Interface of the main page.

The first step is to load the file that the user wants to query into the system, which is done by a simple 'drag and drop' operation. The system will not start unless the file is first uploaded, and will display an alert containing some important disclaimers, in order to warn the user against improper use of this system, or against certain problems that might arise from the operation of the AI. For demonstration, will be uploaded and shown a file retrieved online containing some simple qualitative and quantitative data from various companies. For each 10,000 total organizations, the file contains characteristics such as Country, Website, Description, Foundation Date, Industry or Number of employees. The source of data is an online community (GitHub, n.d.)[14]. Once the system has picked up the file, which will contain quantitative or qualitative data structured in xls, xlsx or csv format, it will display one of two interfaces (by default, the one for querying via text, called the “Assistente Intelligente” or “Intelligent Assistant”) and the user will be able to query the data:

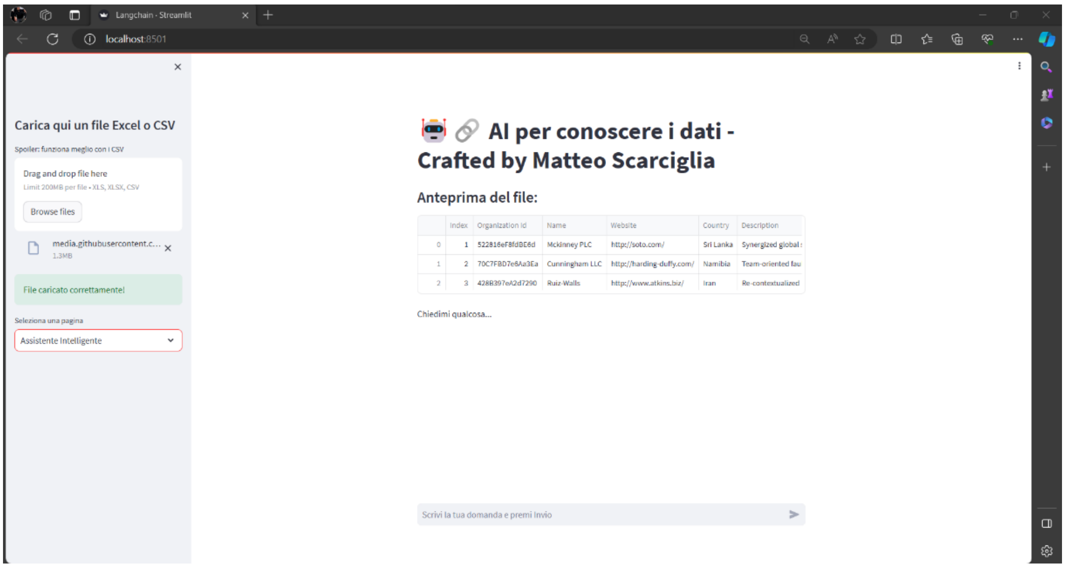


Figure 9 - Main Page ("Assistente Intelligente" or "Intelligent Assistant") once the file is uploaded.

[14] The GitHub repository where the dataset come from: https://github.com/datablist/sample-csv-files.

At this point, a preview of the uploaded file is shown, while the left-hand side shows the file name with the possibility of removing it, the section chosen together with the ribbon that allows you to change the page, and any uploading errors arising from the file, such as a bad date format, not compatible with the query system, or other errors. Here the users can start to submit whatever; then, the prompt is sent to the model, which starts to be trained on data and being able to yield an answer.

In the back-end, various training and response search processes start, resulting in the development of an actual thought, which the developer (but the user, at least in this version, is not) is able to visualise at the command line in the prompt. Below is a demonstration:

Figure 10 - A user-launched prompt, which requests the names of the organisations contained in the file in the first 10 lines.

The request contains 0 to 9 because Python starts counting from 0 inclusive, such that observation number 0 of the list (column) containing all the names is the first.

Meanwhile in the back-end:

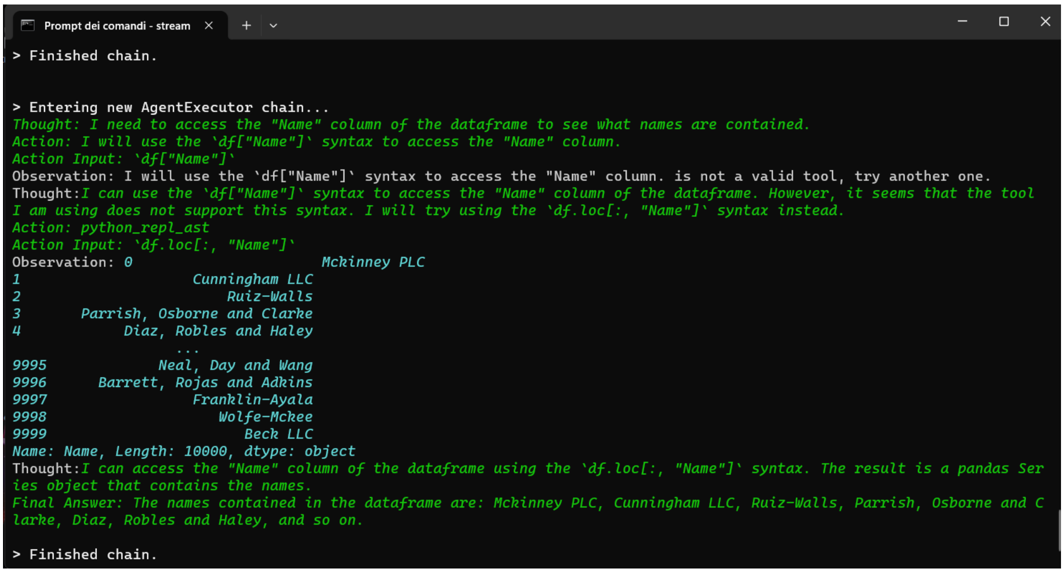


Figure 11 - The AI Agent after being trained on data, thinking.

What is remarkable is the way the Agent thinks about the answer, showing the same reasoning setting as humans. The efficacy of the answers will be discussed later. In addition to the first page, there is a second section dedicated to graphical statistical representations, called “Dashboard”; is also shown here a preview of the uploaded file, as in the previous page.

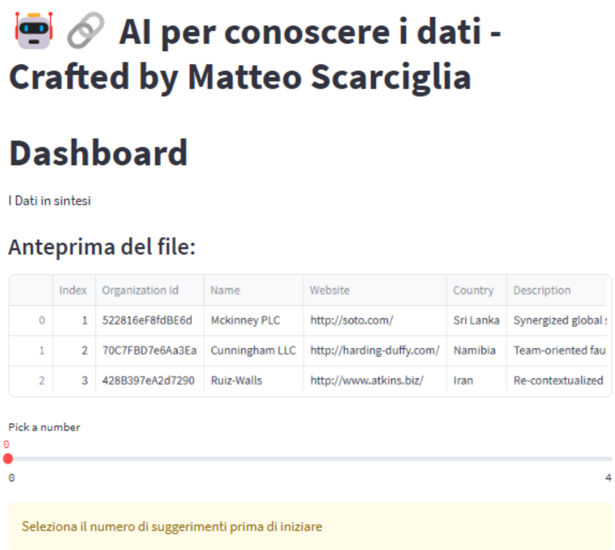


Figure 12 - Main Page of "Dashboard" section.

The first input user has to send is a number, which indicates the number of "graphical tips" the AI Agent has to show. Trying to achieve the objective of increase the Business Intelligence alphabetization of all the actors involved in business, in fact, a system like this can highly help brainstorming among people in all the decision-making, governance and management centre.

So, once the user has chosen the number of tips the AI has to yield to, the Agent starts another thinking process like the previous one which the model is trained on, then the data are analysed and lastly the more efficacy statistical and significative representations are automatically plotted.

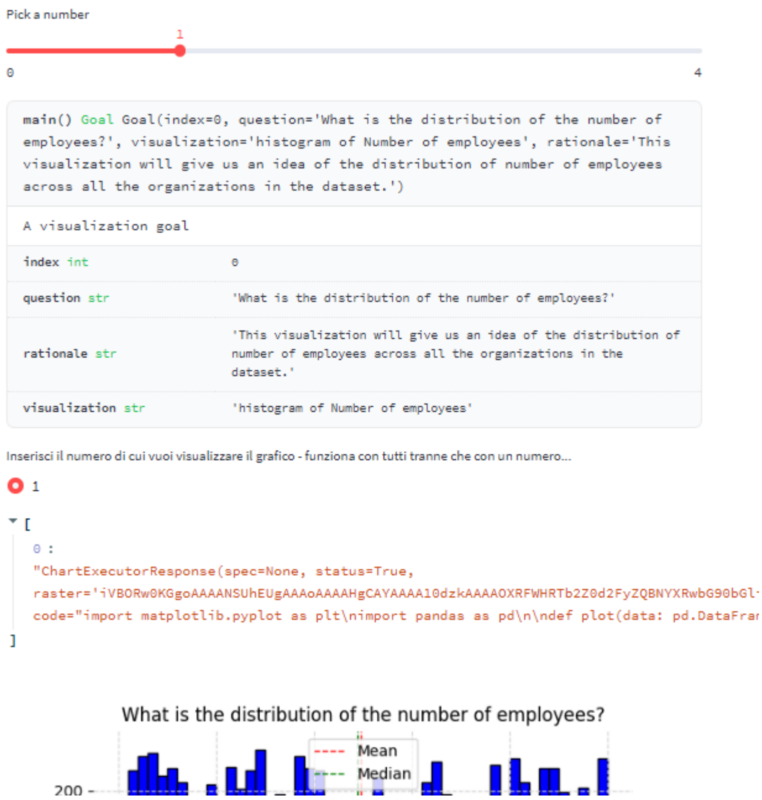


Figure 13 – The results derived from the Agent once the user chooses the number.

At this point, all visualisation tips that the AI considers useful and representative are shown, which are generated according to significance. Settings such as the number of recommendations to be shown are modifiable during programme development.

Given the highly rudimentary nature of the application and the fact that, for the purposes of assessing its proper functioning also in the back-end, some technical characters are shown that do not enrich the final user experience, but serve the programmer to understand what is happening behind it. In this case, a mirror containing the Goal, i.e. the response list object containing the LLM calculation result, is shown on the screen. In this list are the parameters that will be fed to the library's graphics engine, which will take care of displaying the graphical representation on the screen. The variables in the list are *index*, which represents a technical parameter of little significance here; *question*, i.e. the question that the AI has given birth to after the thought process and which will subsequently be editable by the user via a prompt; *visualization*, containing the prompt; *rationale*, containing an intuitive description of the graph, explaining on the one hand how the Agent has deemed that representation meaningful, and on the other hand how it is meaningful to the user. Following, the plot resulting from this process:

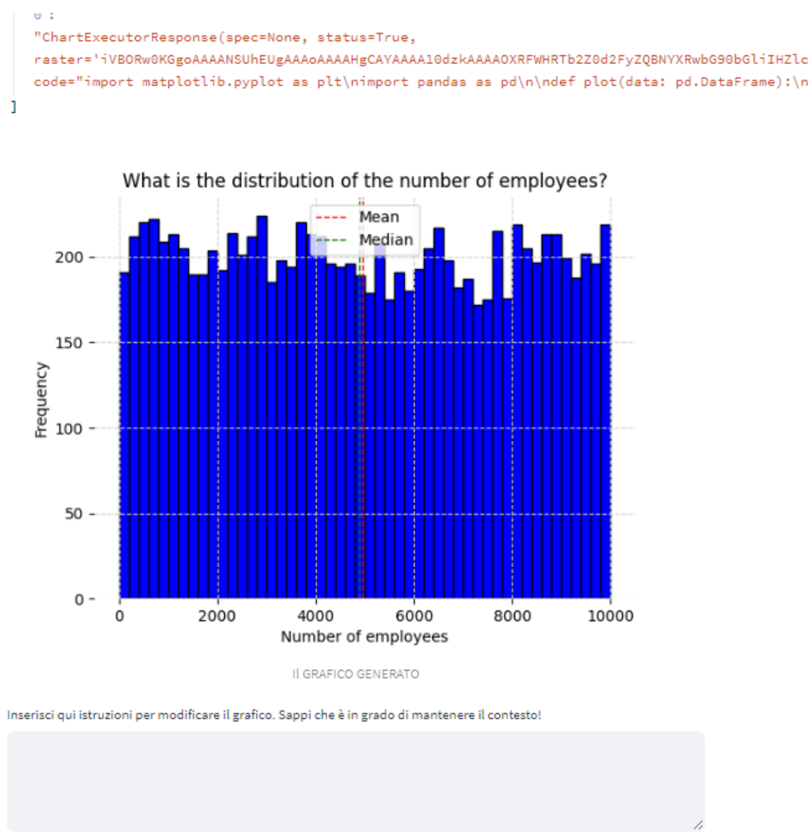


Figure 14 - Histogram of Number of employees showing the distribution of the number of employees, artificially generated.

Once the first plot has been generated, the possibility opens for the user to request graphical changes directly via a prompt. This prompting process will modify the 'visualisation' variable of the list above, resulting in a new instruction for the agent. Below is an example of a change requested by the user, in which a request is made to change the language of the graph from the source language to Italian, to make a dive-in in the categories by increasing the size (frequency) of the y-axis to 50 and then, demonstrating how the Agent is able to process any language, a request to change the colour of the vertical bars from blue to orange:

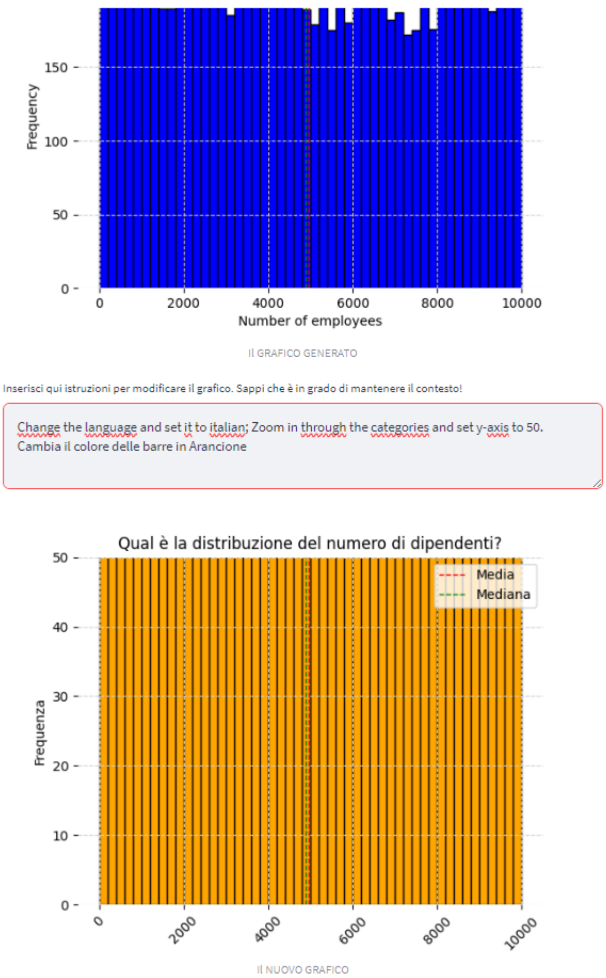


Figure 15 - User's prompt to Agent, where an edit of the graphical representation is asked.

The power of the Agent's logical reasoning, combined with the power of the underlying graphics engine, allows not only for changes to the graph that the Agent is confident in recommending to the user, and which is simpler in appearance, but also to recreate and completely overturn the starting graph, requiring the display of new variables through the most diverse graphical tools, in a totally customisable and absolutely easy way. The following is an example, in which the user requested an ex-novo graphic display, containing information that had nothing to do with the previous one, in which one can perceive the elasticity of the Agent:

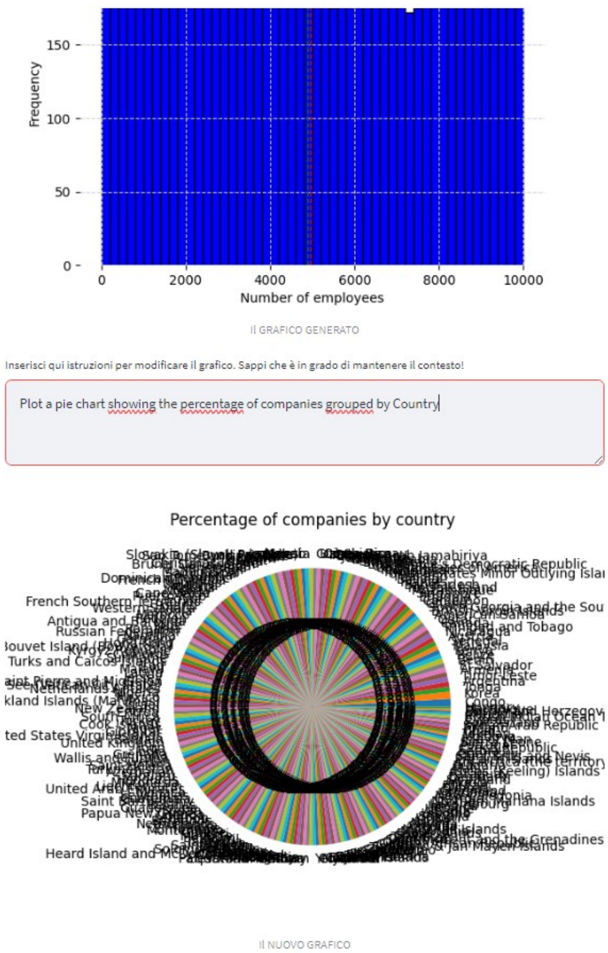


Figure 16 - Graphical representation revisited totally in the typology and information shown.

Clearly, the significance of the graph is not of primary importance here, considering the mere illustrative nature of the prompt; what we wish to demonstrate is merely the way in which, with extreme agility, one can request any change to the graph, thus eliminating any manual creation activity and thus saving time and resources, which can be employed to carry out activities with higher added value, such as in decision-making or management processes. Another demonstration:

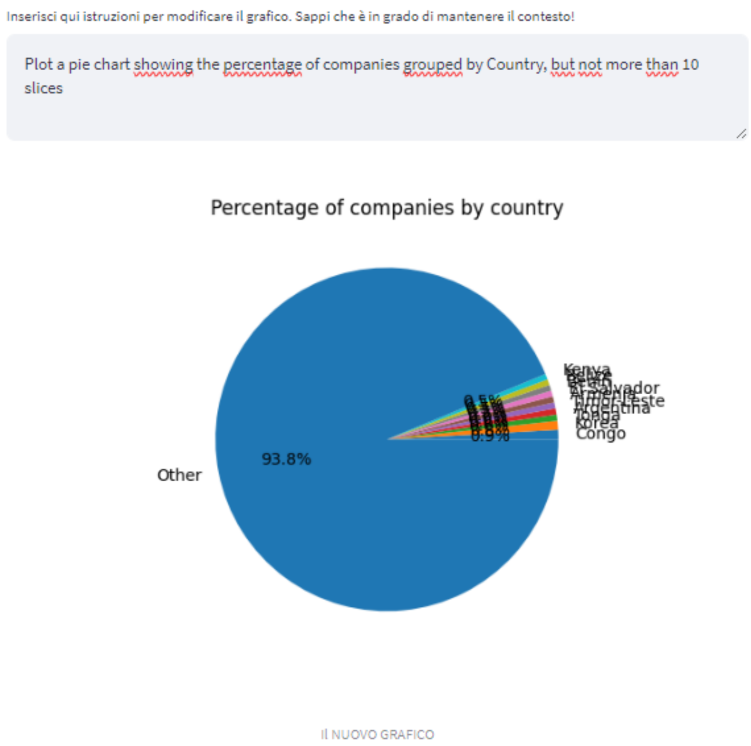


Figure 17 - Another demonstration of graphical editing asked by user.

However, any requested changes occur within the second graph (generated when the user prompt a text requesting an edit) and no older versions than the latest are retained, just as in the case of the smart assistant. This happens in this application due to the poor level of engineering of the back-end, which should certainly be the subject of a more structured study and development in order to easily manage a real conversation and store questions and answers, work that would require the involvement of more verticalized professional figures such as a software developer; in fact, in this context the level of engineering and production is not evaluated as the primary guideline but what is only considered is the ease of construction by a non-expert, and the ease of use perceivable by the user in the face of a very high value information that a system of this type can bring to every organizational level. Having said this, the fact remains that this constitutes, at a conceptual level and not only, an excellent starting point towards which to focus studies on new ways of operating in companies, which can, using increasingly transversal professional figures, access new ways of understanding knowledge thanks to the potential of AI on a side and the ease of implementation that is allowed today on the other. Just as happened in the past times, where the society did not work following the economic scheme of comparative advantage, an economist, for example, now can be a little programmer, an analyst and a scientist at the same time, with advantages in terms of costs, lower information losses, and more empowerment due to the higher level of company's information provided to all actors.

2.1) Libraries and Frameworks for Web-App Development, AI, and Data Analysis

The Python programming language was chosen for a number of reasons, which generally revolve around the fact that it is very easy to learn; even, according to a geek's blog, in just 30 hours[15] it is possible to learn everything about Python, i.e. the main functions, the objects that make it up, how to create a function or even a module. What has just been said, in fact, has made Python one of the most popular, if not the first, among JavaScript, Java, C++ or HTML. This is basically because its ease of use, due to the fact that, unlike the aforementioned programming languages, it does not require too stringent rules of identification or declaration of objects in memory, coupled with the fact that it is totally open source, have consolidated since its birth (the first version of Python, 0.9.0, was released in 1991 by Guido Van Rossum) a very varied community of experts, larger than others.

In this world, each capstone consists of a so-called *library*, i.e. a set of modules and packages that provide pre-written code to perform a variety of predefined tasks. This code, consisting of functions, classes, variables, and calls to functions or objects contained, in turn, in other modules, is called up to develop and implement

[15] More information available here: https://www.a-sapiens.it/informatica/risorse/perche-come-imparare-python/#:~:text=Con%20una%20durata%20complessiva%20di,tutto%20quello%20che%20riguarda%20Python.

common functionality in the programme in which it is used without having to write the code from scratch. Thus, by calling up a library in a single line of code, it is possible to have access to an infinity of functions, so that even the inexperienced person can access powerful tools created by more experienced programmers.

Python libraries can range from small and vertical in one scope, such as those for web scraping, mathematical analysis, or data analysis, to larger and more complex frameworks such as Django, for web development. The libraries used by this program for its operation are varied, therefore constituting an opportunity to provide excellent examples.

The libraries that were used for the development of this application, divided by sector, are those shown below:

```python

# Librerie per AI
import openai

# Prova di LIDA
from lida import Manager, TextGenerationConfig, llm

# Librerie di Langchain
from langchain.chat_models import ChatOpenAI
from langchain.agents import create_pandas_dataframe_agent
from langchain.agents import create_csv_agent
from langchain.llms import OpenAI

# Librerie per Data Analysis
import pandas as pd
import matplotlib.pyplot as plt
import tkinter
import plotly.express as px

#Librerie per Web App ed altri componenti
import streamlit as st
from streamlit_echarts import st_echarts

```

Figure 168 - The instance of the script, where all libraries are recalled.

Among these, the most diffused and also most used in the whole *script* are those dedicated to Data Analysis, where libraries for data visualization, analysis and mathematical object manipulation are included. Starting from *Pandas*[16], this is a power, flexible and easy to use library fully dedicated to data analysis, specialized in treating structured sources of data in the form of an Excel's table; it introduces two new data structures to Python, a bidimensional structure similar to a table called *DataFrame*, and a unidimensional structure similar to a column of a table, called *Serie*. Additionally, with Pandas it is possible to execute a variety of operations on data, like filtering, selecting, editing, join or group operations (SQL like), missing data handling, and so on. What makes Pandas smart and scalable is the capacity of importing and exporting data in a very large variety of format, such as CSV, Excel, SQL, JSON, and a lot more which can be even more extended, recalling other libraries, which is the other diamond tip of Pandas. Another among the most important libraries dedicated to data analysis operations is *Matplotlib.* This is one of the most popular and powerful data visualization libraries, providing a very flexible way to build various graphical representations and plots. Matplotlib[17] is characterized for its easy of use, which makes it beginner friendly thanks to a simple interface which allow to create graphical representation writing only a few rows of code. Additionally, it is one of the most versatile libraries due to its capacity of integration with libraries such as Pandas or NumPy[18], which is the most powerful library for mathematical analysis operations, and lastly thanks to its extensibility towards other under-libraries like Seaborn[19] (very popular for creating advanced statistical representations or mpl_toolkits, used for additional functionalities such as 3D graphs). The other libraries shown in Figure 18 under the Data Analysis section are surely important, but their existence here is mainly for completion of other libraries, which require these in the back-end as dependencies.

Among other libraries, those that truly makes a disruption are that of Artificial Intelligence *OpenAi*, released by the homonymous organization which allowed its

[16] More information about Pandas available on: https://pandas.pydata.org/.
[17] More information about Matplotlib available on: https://matplotlib.org/.
[18] More information about NumPy available on: https://numpy.org/.
[19] More information about Seaborn available on: https://seaborn.pydata.org/.

existence, and *Streamlit*[20], another very powerful and innovative library that has allowed low-code development of a system like the one examined here. The OpenAi library is a composite of various Application Programming Interfaces (API) and other tools developed to provide access to the user to the most advanced Artificial Intelligence models, especially in those dedicated to Natural Language Processing (NLP) and Machine Learning. The library provide an easy access to its APIs through just a few of code and users can employ Generative Pre-trained Transformer (GPT) models for generating text, answering questions, reassuming documents or translating languages, or either other models like *DALL-E*[21] which can generate images starting from textual description provided by user or *Codex*[22], which is a variety of *GPT-3* model declined to assist users while programming, able to understand chunks of code and generate whole algorithms from scratch. Looking at Streamlit instead, it is a powerful open-source framework build to create applications dedicated to data analysis and visualization quickly, became diffused among data scientists, analysts and other domains users due to its easy of use. Its functioning is simple also, because it is enough to write just some code which is simultaneously transformed in a Graphical User interactive Interface directly on a web-based application. Its advantages and easy of use reside on its high compatibility with a lot of third-party components (widgets, sliders, text box and other) which allow final user to interact directly with data, which are handled using Pandas or Numpy, and visualized through Matplotlib or Plotly.

Moreover, Streamlit build applications can be easily shared and deployed thanks to some dedicated platforms, and its open-source-ness has formed a very large community and a lot of resources available such as tutorial, documentation or forums.

---

[20] More information about Streamlit available on: https://streamlit.io/.
[21] More information about DALL-E available on: https://openai.com/research/dall-e.
[22] More information about Codex available on: https://openai.com/blog/openai-codex.

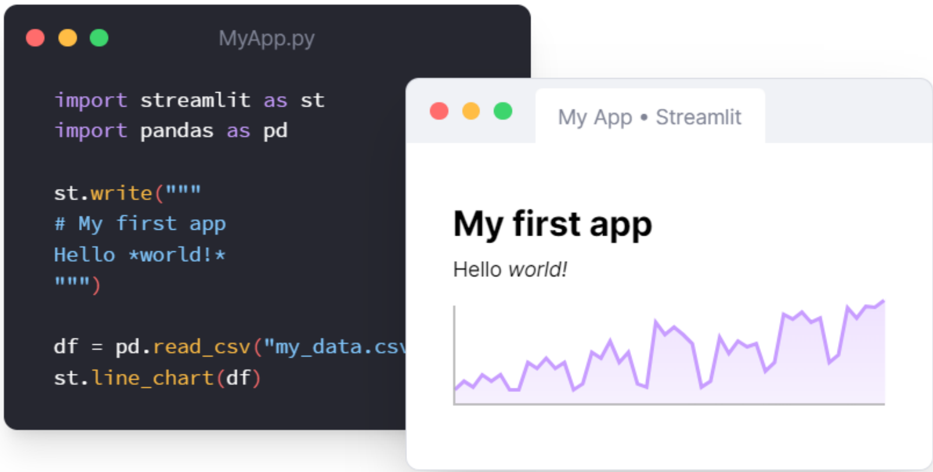


Figure 17 - Ease of use shown on the home page of Streamlit website.

Although Streamlit is an excellent tool due to its rapidity and ease of development, shows some limitations which reduce its suitability on large scale projects, or in those which an high presence of customized elements is required; mainly, the “out-of-the-box” nature of its elements does not offer the same level of control obtained when every element is developed manually using other tools like HTML, CSS or JacaScript. This can bring less scalability and flexibility compared to the complexity sometimes required in some projects; other libraries more suitable for this scope can be *Django* or *Flask*, which offer an higher control on user interface and especially on the application’s architecture, enhancing stability, safety and scalability making them preferable for more robust web-applications.

2.2) Different LLMs comparison and API integration in the system

The rise of LLMs revolutionized the business landscape, compelling companies across various sectors to innovate and adapt to stay competitive. This shift is marked by substantial investments in research and development, recruitment and training of AI-specialized talent, as well as strategic collaborations with industry leaders like OpenAI and Google DeepMind. Businesses are leveraging these advanced technologies to develop new products and services, enhance operational efficiency, and address challenges in ethics and regulatory compliance. The adoption and integration of LLMs are thus transforming how companies operate, communicate, and innovate, marking a significant shift in global business strategies. Some LLMs, reported here due to their balance to be easy to implement yet composed of the biggest number of parameters, are:

- GPT-4[23] (OpenAI);
- PaLM 2[24] (Google);
- Llama 2[25] (Meta);
- Gemini[26] (Google DeepMind).

Starting from GPT-4, this LLM has the characteristic of being trained on 175 billion parameters, ten times as many as its predecessor, GPT-3.5. This model immediately presented itself as highly capable of generating realistic and creative text, writing code or answering questions while maintaining perfect context. It is currently among the model with the largest number of parameters and relies on incredible computing power.

Then PaLM 2 (Pathways Language Model 2) is another important model, developed by Google. This has a parameter count of 540 billion, competing not only with GPT-4 but also with other large models. Among the strengths of PaLM 2 is its distinct ability to go beyond simple text generation, as it can reason about facts and relationships, solve problems and even learn new concepts. In any case, it is perfectly suited to translation or content creation tasks, and thanks to Google's proprietary 'Pathways' architecture, it can easily be adapted to new tasks and domains.

---

[23] More information about GPT-4 of OpenAI available on: https://openai.com/gpt-4.
[24] More information about PaLM 2 of Google available on: https://ai.google/discover/palm2.
[25] More information about Llama 2 of Meta available on: https://llama.meta.com/llama2.
[26] More information about Gemini of Google DeepMind available on: https://deepmind.google/technologies/gemini/#introduction.

Meta (formerly Facebook) is also carving out its own space within this sphere with its Llama 2 model, publicly released in July 2023. Released in three versions of 7, 13 and 70 billion parameters, it is positioned as a mid-range model in terms of size, offering a good compromise between efficiency and capacity. Unlike competing models such as OpenAI's or Google's, it is not an end-user application, but on the contrary has been open-sourced to all developers, and thus designed for professional and academic use. Among the strengths of Llama 2 are its efficiency, versatility and easy accessibility, as it can be applied to solve various tasks such as text generation, translation, text summarisation, answering questions or generating other content.

Lastly Gemini is one of largest LLM developed by Google DeepMind, recently released. It is based on “Transformer” and “Sparse Attention”[27] (Child R., 2019) neural networks, which allows this model to have multimodal capabilities enabling advanced skills in natural language understanding and thus in the generation of realistic and coherent text, in the translation of languages taking into account culture and context, or in the writing of high-quality creative content. This type of multimodal model, characterised in three categories - Ultra, Pro and Nano - is capable of outperforming existing benchmarks in 30 out of 32 cases, setting new standards especially in the areas of multimodal performance and cross-modal reasoning. As the Google DeepMind technical report[28] shows, Gemini Ultra's reasoning ability outperforms other models, including GPT-4 and PaLM 2, in reasoning tasks and contextual understanding, also showing a marked ability to retrieve accurate information from even extended contexts. Gemini, for the purpose under discussion here, would therefore find perfect application; however, both for reasons of ease of implementation and less established use by a large number of users, the model that was used was GPT-4. In any case, a comparison between the

---

[27] 2019, Child R., Gray S., Radford A., Sutskever I., “Generating Long Sequences with Sparse Transformers”, Cornell University, https://arxiv.org/abs/1904.10509v1 of 23 April 2019.
[28] Technical Report of Gemini available here: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

two models in the performance of some more or less complex tasks will be carried out in Table 1 below, analysing their score against the relevant benchmark.

2.3) Capabilities of the Agent – Advantages, limitation, and prompt-efficacy

The performance of an LLM is measured adopting various qualitative or quantitative methods both based on specific benchmark. Some quantitative measures could be *Accuracy*, which measures how many times the answers yielded are correct out of a defined set of test data, or *Perplexity*, which is used to evaluate how well the model is able to predict the next word in a period. Although highly relevant, for more specific and technical information, please refer to specific performance evaluation works, which can provide an objective and proven analysis of the ability of an LLM.

Here, what will be the subject of a somewhat more in-depth evaluation are the qualitative evaluations of a model, which focus on the evaluation and human perception of the generated responses, and which are based on criteria such as relevance or coherence. Some criteria that would be interesting to analyse concern aspects of the experience of using an LLM, such as the speed of inference (the speed with which a model can generate answers) or the use of resources (how many computational resources are needed to train and run the model), or even the fairness and bias contained in the responses generated or robustness, as the ability to manage unexpected inputs without excessively degrading performance.

Google DeepMind, showing the performance of its leading model (Google DeepMind, n.d.)[29] on its website compared some capabilities of its model with the

[29] More information on Gemini available here: https://deepmind.google/technologies/gemini/.

GPT-4 model of the competitor OpenAI. Reporting what is analysed by Google, the Capabilities measured, and the relative benchmarks are the following:

Capability	Benchmark	Description	**Gemini Ultra**	**GPT-4**
General	MMLU	*Representation of questions in 57 subjects (incl. STEM, humanities, and others)*	90.00%	86.40%
Reasoning	Big-Bench Hard	*Diverse set of challenging tasks requiring multi-step reasoning*	83.60%	83.10%
	DROP	*Reading comprehension (F1 Score)*	82.4	80.93
	HellaSwag	*Commonsense reasoning for everyday tasks*	87.80%	95.30%
Math	GSM8K	*Basic arithmetic manipulations (incl. Grade School math problems)*	94.40%	92.00%
	MATH	*Challenging math problems (incl. algebra, geometry, pre-calculus, and others)*	53.20%	52.90%
Code	HumanEval	*Python code generation*	74.40%	67.00%
	Natural2Code	*Python code generation. New held out dataset HumanEval-like, not leaked on the web*	74.90%	73.90%

Table 1 - Google's Gemini Ultra LLM compared to GPT - 4 by OpenAI (source: Google DeepMind).

Henceforth, in the evaluation of the model used to “motorise” the Agent in this project, this capabilities framework will be, approximately at least, used, in order to make the evaluation of effectiveness more practical for the purpose for which it is used here and referring to more specific literature for further information. Here, given the purposes of this system, a series of questions were posed to assess, basically, the ability to understand the (business) context, to solve more or less complex operations and problems of text classification, translation or mathematical operations, the 'emotional quotient', to assess the human aspect in the response (both in being able to identify with the user and in the resulting confidence, which is directly proportional to the emotional quotient) and finally the Prompt Engineering work required, i.e. the effort the user must make to obtain a consistent response.

Following, some tests:

What are the names of the top 10 companies?

The dataset provided contains information about various companies, including their names. To answer the question "What are the names of the top 10 companies?", we can simply extract the names of the companies from the "Name" column in the dataset.

Here are the names of the top 10 companies:

1. Mckinney PLC
2. Cunningham LLC
3. Ruiz-Walls
4. Parrish, Osborne and Clarke
5. Diaz, Robles and Haley
6. Smith, Ramos and Barnes
7. Jefferson, Ross and White
8. Johnson, Clark and Foster
9. Smith, Johnson and Brown
10. Brown and Sons

Please note that these names are based solely on the information provided in the dataset and may not reflect the actual rankings of the companies.

Figure 20 - The first question submitted to the Agent.

Analysing the content of the answer, nothing of extremely relevant is noticed, as the Agent answered perfectly showing the top 10 companies as listed in the dataset; noticeable is the process which led to the answer, which is explained just after the question instead. Continuing with a slightly more complex prompt, a demonstration of the average value prompt in the 'Number of employees' column is given, with the aim of testing the ability to perform basic statistical calculations:

What is the average value of the number of employees? Provide me the value

To determine the average value of the number of employees, we need to analyze the "Number of employees" column in the dataset. We can do this by calculating the mean value of this column.

Here is the calculated average value:

Average number of employees: 5322.407

Figure 21 - The second question submitted to the Agent, more difficult.

Here too, as before, the LLM does quite well, providing a rationale and the solution found. Now, a more complex prompt exercise is provided, which attempt is to evaluate the capacity of calculating more complex statistics, such as interquartile range and dispersion:

What is the interquartile range (IQR) for the Number of employees column? And what does this value indicate in terms of data dispersion? Provide me the IQR and the associated numerical solutions

To calculate the interquartile range (IQR) for the "Number of employees" column, we need to find the first quartile (Q1) and the third quartile (Q3).

First, we need to sort the values in the "Number of employees" column in ascending order. Then, we need to identify the position of the first quartile (Q1) and the third quartile (Q3).

Q1 is the median of the lower half of the dataset, and Q3 is the median of the upper half of the dataset. To find the position of Q1 and Q3, we can use the formula:

Q1 = (n+1)/4 Q3 = 3(n+1)/4

Where n is the total number of observations.

Once we have the positions of Q1 and Q3, we can find the corresponding values in the dataset. The interquartile range (IQR) is then calculated as:

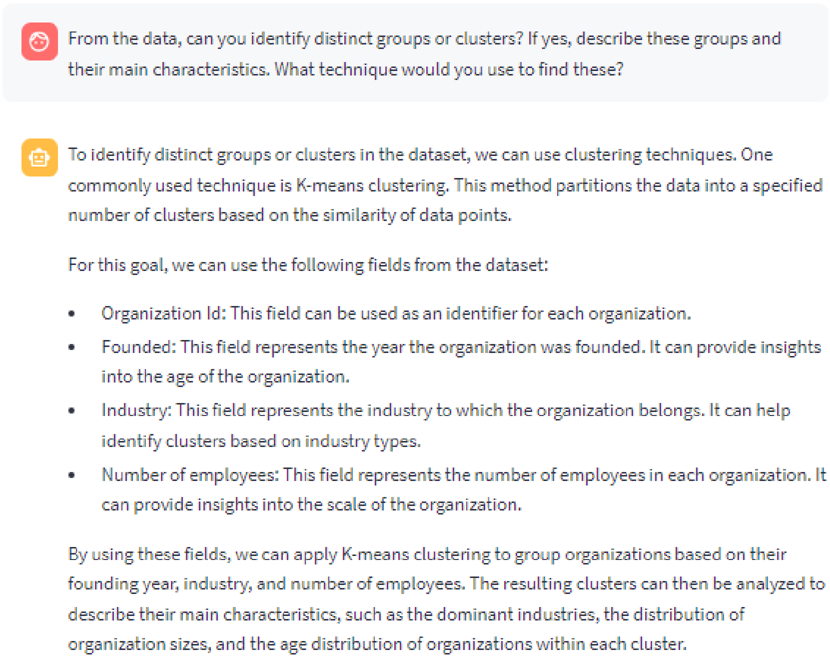
IQR = Q3 - Q1

Let's calculate the IQR for the "Number of employees" column using the given dataset.

Figure 22 - The third submitted question, even more difficult.

It turns out that the model is able to understand the context and provide a clear definition of what is required of it, presented in such a way that it can be easily understood by the reader. However, despite several attempts at prompting with the same objective but posed in more or less different ways, the response, which demonstrates a high level of intelligence through its ability to break down the IQR calculation process into logical and comprehensible steps, merely provides methodical guidance and not the practical example applied to the data in question.

Therefore, the answer is only partially satisfactory as it does not complete this task with an actual calculation. Another demonstration:



From the data, can you identify distinct groups or clusters? If yes, describe these groups and their main characteristics. What technique would you use to find these?

To identify distinct groups or clusters in the dataset, we can use clustering techniques. One commonly used technique is K-means clustering. This method partitions the data into a specified number of clusters based on the similarity of data points.

For this goal, we can use the following fields from the dataset:

- Organization Id: This field can be used as an identifier for each organization.
- Founded: This field represents the year the organization was founded. It can provide insights into the age of the organization.
- Industry: This field represents the industry to which the organization belongs. It can help identify clusters based on industry types.
- Number of employees: This field represents the number of employees in each organization. It can provide insights into the scale of the organization.

By using these fields, we can apply K-means clustering to group organizations based on their founding year, industry, and number of employees. The resulting clusters can then be analyzed to describe their main characteristics, such as the dominant industries, the distribution of organization sizes, and the age distribution of organizations within each cluster.

Figure 23 - The fourth question submitted to the Agent, asking for clusters.

The answer illustrates the application of a methodical approach to the problem, applying the cluster identification through the k-means algorithm. What the Agent have done is a primarily activity of research looking for what algorithm would be, in its opinion, the most appropriate to solve this problem; again, although it demonstrates a good understanding of the context, the requirement and also the method to be applied, the explanation lacks practical examples and a fuller discussion of the expected results of the analysis. The approach is well explained but should be enriched by a concrete demonstration of the implementation and analysis of the resulting clusters, including insights and distinctive characteristics of these clusters.

More serious, however, is the use of the variable "Organisation ID", which is totally inappropriate to be a common element of clusters as it represents a unique identifier for each company. In any case, expecting such a system to be perfectly capable of understanding the problem and, above all, of providing a response as varied and complex as that required in the most difficult prompts is unrealistic, given the still immature state of evolution of the underlying technology. The solution mode, which would allow answers to more complicated problems, would maybe be better with the implementation of certain steps in the prompting phase of the back-end, which could allow to undertake different solutions, or requesting as a response some programming code applied directly to the data, which can in turn be fed back to the AI requesting only the response in exchange, both resulting in better results.

However, there is still a long way to go and Generative AI, basing its evolution and the underlying operating mechanism on reiterated automatic learning processes, still has a lot to learn from its own mistakes.

Chapter 3) Impact of a LLM based AI Agent in Business Operations – A new Business Model Taxonomy

Generative AI, thanks to its high innovativeness due to the great attention paid by a large number of researchers and the technological possibilities deriving from it, in terms of customisation of uses in various combinations, is already revolutionising the way companies operate, bringing significant improvements in terms of efficiency and innovation, allowing Business Intelligence to be done in a totally new way. We are therefore witnessing the presence of a veritable new figure, a super consultant capable of providing more or less complex answers based on text or graphical support, capable of generating new knowledge more quickly and efficiently[30] (Jermakowicz E., 2023). LLMs are capable of generating several kind of contents, such as articles, reports, creative writing, or more complex activities such as sentiment analysis, question answering and data analysis, due to their focus on language pattern recognition, perception and generation. These allowed Generative AI to introduce new paradigms of interaction, operativity and strategy in companies, especially considering his capability of listen and interpreting human thinking, and acting best to solve every problem human can think of. To explain the taxonomy of AI-driven business, Osterwalder's Business Model Canvas (Osterwalder, 2005) will be used, a framework that studies the various components of a business model, explaining how they fit together and how they are able to generate value. Osterwalder's model consists of nine blocks, which are:

- *Customer Segments*: in this area are explained who the customers targeted by the business model are, including their preferences, so they are clustered by some characteristics, whether they belong to the Mass market, a Niche market or a Multi-Sided platform/market;
- *Value Propositions*: the value proposition address the way in which the business build value to the customers, providing new elements in the

[30] Jermakowicz E. K., 2023, "The Coming Transformative Impact of Large Language Models and Artificial Intelligence on Global Business and Education", St. John's University, https://scholar.stjohns.edu/jga/vol4/iss2/3/ of 21 December 2023.

offering such as newness, design, low price, brand and generally what distinguishes the company from its competitors;

- *Customer Relationships*: describes the types of relationships a company establishes with specific customer segments, exploring how a company interacts with its customers trying to enhance, for example customer loyalty or retention, or generally how the business will get new customers;
- *Channels*: refer to how a company communicates with and reaches its customer segments to deliver the value proposition. They are important touchpoints and can be more or less direct (depending on how human's touch is involved in) and more or less customized, based on target;
- *Key Activities*: these are the actions a company must take to operate successfully in the market, since are critical to making the business model work. Key activities include production activities and generally all the activities belonging to the primary or supporting Michael Porter's (Porter, 1985) pool of activities (Tarver E., 2021)[31];
- *Key Resources*: here belong all the assets required to offer and deliver the previously mentioned elements (value proposition, channels, customer relationships, revenue streams) and can be physical, financial, intellectual, or human;
- *Key Partners*: this area identifies what are the key alliances involved in building value for customers, including suppliers, partners or other kind of alliances or collaborations that can help a company optimize its operations, reduce risks or acquire resources more conveniently;
- *Revenue Streams*: here there are the sources from which a company generates revenues, resulting, for example, from one-time customer payments, recurring payments or after-sales services to the customer in exchange for payments;

[31] Tarver, E., 2021, "What Are the Primary Activities of Michael Porter's Value Chain?", Investopedia, https://www.investopedia.com/ask/answers/050115/what-are-primary-activities-michael-porters-value-chain.asp of 01 September 2021.

- *Cost Structure*: lastly this area describes all the cost the company has to sustain in order to generate the value offered, including all the facilities, goods, and any other manufacturing and delivery related costs. This area defines majorly, besides of the value proposition, the settings of a company in terms of branding, since costs, within the business, are used as first driver while deciding prices.

Analysing the business transformative impact of the LLMs usage, according to the research "Towards a Taxonomy of Large Language Model based Business Model Transformation" (Wulf J.) of Zurich University of Applied Sciences[32], data showed four archetypes of Business Model Redefinition: 1) Customer renewed benefits, 2) New channels of communication and sales, 3) New capabilities of doing Business Process Automation, 4) Improved use of information.

Starting from the first point, regarding the new kind of customer benefits, researchers demonstrated through data and empirical examples that LLMs are used in this area to bring a new customer-centred experience while bringing value, being installed in applications for providing personalized information to users related to the service for *Personal Assistance*; sometimes LLMs showed their capacity to support user's learning path in online courses, especially in correcting exercises, directly improving results acting as a *Teacher*; other companies use LLMs to provide a more interactive experience with their products, such as *Speech Interaction* integrated in some kind of virtual assistant; lastly, some textual components have found online, showing how well can be LLMs applied to write product descriptions or other *Content Generation*.

Continuing, the way LLMs allowed new sales and communications channels is, for example, bringing users personalized information when a question has been asked, such as a product's characteristic which is not yet publicly available on a website

[32] Meierhofer J., Wulf J., 2023, "Towards a Taxonomy of Large Language Model based Business Model Transformation", Zurich University of Applied Sciences, https://arxiv.org/ftp/arxiv/papers/2311/2311.05288.pdf of 09 November 2023.

thus supporting *Presales Automation*, or *Customer Service Automation* systems able to handle customer inquires in after sales phases has also been found;

Again, in the domain of Business Process Automation, LLMs has brought enhancing in the automation of information-involving business processes, such in the case of *Front Office Process Automation*, which encompasses customers in the way (but not only) as previously demonstrated, automating company-internal tasks for data handling and transferring optimizing the *Back Office Process Automation*, or either in *Software Development Automation*, where LLMs definitely and irreversibly changed the way of coding, creating whole chunks of more or less complex code writing just a prompt.

Ultimately, LLMs have changed the way of information usage in companies, allowing, as demonstrated in the previously showed application, which is the perfect demonstration of AI-powered Software Development also, a strong enhancement of utilization of information, both in terms of quality and quantity. Indeed, some major institutions such as Morgan Stanley[33] (Son H., 2023), which build a LLM-powered assistant for financial advisors helping them in *Information and Knowledge Management* or *Information Extraction.*

A research paper[34] (Brynjolfsson E., 2023) demonstrated whether individual level changes in communication lead low- and high-skill worker to sound more alike. Following, directly from this paper, the text similarity between high- and low-skill agents at specific moments, separately for workers with access to AI Agent (blue dots) and without (red diamonds), showing that the difference on textual production between high-productivity and skilled workers low-productivity and skilled

---

[33] Son H., 2023, "Morgan Stanley kicks off generative AI era on Wall Street with assistant for financial advisors", CNBC, https://www.cnbc.com/2023/09/18/morgan-stanley-chatgpt-financial-advisors.html of 18 September 2023.
[34] Brynjolfsson E., Li D., Raymond L., 2023, "Generative AI At Work", NBER, https://www.nber.org/system/files/working_papers/w31161/w31161.pdf of April 2023.

workers is 0.55 among workers which are not using Generative AI, and how the situation changed:

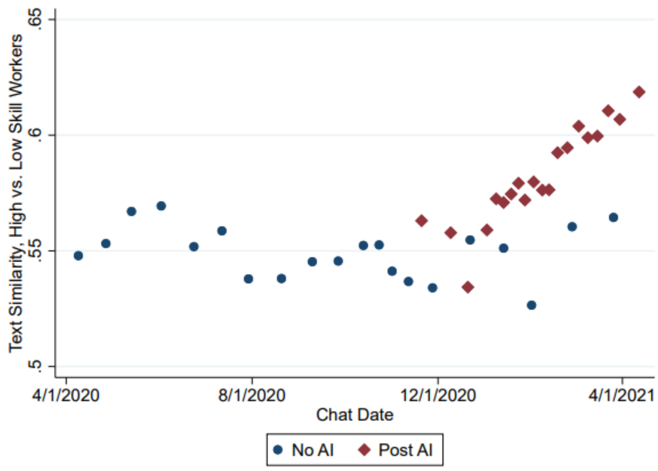


Figure 24 - Text Similarity between High skill and Low skill workers (Source: https://www.nber.org/system/files/working_papers/w31161/w31161.pdf).

What is shown is that with the usage of Generative AI, this difference tends to be smaller, with a text similarity increasing up to 0.67, suggesting that AI can efficiently narrow language gaps. What the research above aims to show is how workers engage generative artificial intelligence, especially workers with fewer skills. In general, it is thus proven that workers who tend to follow the agent's recommendations have much higher returns, especially considering the narrowing of the skills gap. Overall, the penetration of Generative AI in the everyday work in almost the totality of the industries could increase the demand for workers who possess complementary skills such as programming, data analysis, and research.

In general, Generative AI is revolutionizing the business taxonomy radically changing the traditional structures and hierarchies of the market introducing new business models and innovative strategies, making competition more dynamic allowing to a higher number of people to become a part of. This happens thanks to a "democratization" of the knowledge, which facilitates the entrance of new actors in the market culling the gap between high and low competencies, both for individuals and whole companies, in the lower level such as for small companies and at the higher as for bigger; companies are changing the way they approach to the market and to the business model, and this evolution does not only concerns solving specific and limited tasks, but, as demonstrated, can be applied for a lot of innovative applications redefining the competitive landscape.

3.1) Development possibilities of an LLM-powered Information System and Cost-Benefit analysis

In an always changing business environment, characterized by an ever-increasing competition and by the need of optimize the available resources, the LLMs integration represents a new frontier having significant economic implications. Looking through this section, will be explained the cost-benefit analysis derived from the adoption of such technology in organizations, evaluating how LLMs can influence the operative efficiency, productivity and competitivity on the market.

Thus, looking at this kind of Agent as a powerful module or as a whole Information System, a functional analysis of the architecture can be done as happens for a common Information System; the analysis of an LLM-based Agent as an Information System (or an Information System AI-powered) requires an holistic approach which takes into account several aspects, such as Technological Structure, Data Organization, Scalability and Integration, and Data Safety. Every aspect deserves a specific evaluation, which is provided following.

3.2) Architecture and Technological Structure

The architectural scheme of an LLM-based Information System lies in the specific company's needs, depending on the quantity and variety of data to elaborate and to the objective of the system. The more convenient and common architecture which adapt to this kind of system could expect the convergence of all the information, data and request towards a unique elaborative node, as happens for *Centralized*

Architecture[35] (Lavecchia V., n.d.), or can otherwise be based on a distributed elaborative operations among more nodes, as for *Distributed Architecture*. The choose of architecture is made once the needs are defined, since the realization of this kind of system has to take in account its high need for integration of different technological components. Indeed, the choice of infrastructural elements such as the hardware involved in handling, conservation and transfer of information and data, is a critical operation to be done as first. Since the algorithms which allow LLMs work employ heavily Graphical Processing Units (GPU) and TPU (Tensor Processing Units) components, factor as difficult of implementation, maintenance and costs are to be considered before structuring an architecture; also, the Software and developing framework required in order to make a working system have to be perfectly synchronized with hardware and perfectly selected and manipulated to get an efficient and effective usage of resources, and for these reasons a *Cloud-Native Architecture*[36] (AWS, n.d.), perhaps outsourced to a third-part provider, can be a solution that can allow organizations to do not think about hardware maintenance, avoiding having to support high investments in technological too, but anyway beneficing of technological advancement and consumption-based costs.

While functional analysis of how an Information System is managed reveals that, such a model can indeed be totally created and managed in-house by a company, the manage of the LLM component of this system is not easy as for other components or modules, because companies are not currently able to handle such infrastructures and know-how without outsourcing. In this case, the use of the model takes place outside the company's boundaries, with the consequent occurrence of both advantageous and risky situations. When the model is hosted by a third party, it means that any requests are taken and managed by hardware components held by someone else. From the point of view of advantages, this

[35] More information available on: https://vitolavecchia.altervista.org/sistemi-centralizzati-sistemi-distribuiti/.

[36] More information available here: https://aws.amazon.com/it/what-is/cloud-native/#:~:text=The%20cloud%2Dnative%20architecture%20combines,blocks%20of%20cloud%2Dnative%20architecture.

certainly leads the user to avoid all those costs related to the construction, management and maintenance of these infrastructures, which mostly consist of computers or computer components (such as graphics cards) of very high prestige and specifically designed to handle inference tasks. An estimate by TrendForce[37] (Tseng P.K., 2023) showed that in 2020, ChatGPT required around 20,000 GPUs (Graphics Processing Units) to process the data underlying the service's training. The solution OpenAI uses is NVIDIA's A100 card, an accelerator costing around $15,000 per unit. For a company to incur such costs to implement such a system internally would be unthinkable, at least in this day and age. Admittedly, the AI race has only just begun and the market may be in a real bubble at the moment, but it is also true that it would still be better to leave the achievement of economies of scale to specialised third parties, especially considering the costs of utilisation, against the advantages that outsourcing offers, which will be analysed in a later section.

Talking about risks, one of the most important risk companies may face is related to data safety. Generally, wherever data travel from one place to another several risks may configure, such data breach, damaging or destruction. Although today is even easier to handle data transmission and prevent data damaging or destruction, data safety is becoming increasingly a problem looking at how these can be precious in business and how third parties could take advantage of this, unintentionally attracting interest from others. Moreover, once the data safely leave company boundaries the control on these is lost and the only thing to do is to rely on the same provider security systems and their guaranteed maintenance. In any case, the architecture of an information system of this type maintained on premise and some security aspects will be discussed later in this chapter.

- Data Organization:

[37] 2023, Tseng P.K., "TrendForce Says with Cloud Companies Initiating AI Arms Race, GPU Demand from ChatGPT Could Reach 30,000 Chips as It Readies for Commercialization", TrendForce, https://www.trendforce.com/presscenter/news/20230301-11584.html of 01 March 2023.

The organization of data, in terms of what they are, what they do refer to, where they are, where they do travel from, and what is the medium of transmission, assume a crucial role in the Information System working, and consequently on the whole organisation management. Situations of data mala-gestio can significantly vary the quality of the information in output, since that quality depends strongly on data quality, variety and relevance. The first step of the data organization process is to collect data, which consists of all the activities required to identify all the sources where the data are to be collected from, assuring that they are automatically prepared (in an ideal scenario at least) to be handled by a variety of applications of model, formatting them properly (for example, codify data in XML, ready to be transferred across different applications) and automatically transformed. This process became within the *Extract/Transform/Load (ETL)*[38] pool of activity, which is the second step of the data organization process consisting of *Data Extraction*, where data are extracted from the relative origins such as documents, e-mail, csv file from ERP or database and located in a structured or unstructured way, depending on the needs, in destination such as Data Lakes or Data Warehouses; at this stage, the data are to be still considered as raw, since they are only collected and in the most of the circumstances they present errors, incongruences or information which are not pertinent to the phenomenon took in consideration, and the consequent phase starts, which is *Transformation*. During this phase, all the corrupted data such as duplicated records, missed standardization, unneeded sensible information or more generally, errors, must be removed to avoid a bad representation of the reality. This is the most critical moment, because here are applied all the format and rule necessary to complete the elaborative process to come and data are also subjected to treatments like labelling, grouping standardization, integration from other sources, checking and ordering. Moreover, especially in the context of LLM which are particularly sensitive to natural language, every source of data containing propositional information should be purified from language or cultural biases, which is done introducing possibly the

[38] More information available here: https://www.oracle.com/it/integration/what-is-etl/.

most objective source of data, even in other languages if the original is deemed too polarized towards a culture. Once the Transformation phase is finished, the data extracted and properly transformed are ready to reach the destination where they will then be stored, directly processed or transferred, or more properly *Loaded*. In the case of LLMs, the intended destination corresponds to any support capable of being easily processed in the subsequent data extraction phases, such as a data warehouse or database. To date, LLMs are not yet directly integrated with the most common supports, except through service channels activated through specific APIs, which are customized to work with certain technologies in order to formulate data constructs suitable to be data more easily fed to an LLM, as pdf, csv or json.

- Scalability and Integration:

The analysis of the scalability and integration of a LLM-based system requires to consider the way the organization is set respect the probably evolution of technology, and how the governance and management intend to face all the threads and catch all the opportunities. The scalability is a requirement which refers to the capacity for a system to expand or reduce flexibly the required technological resources and thus to adapt itself to a changing environment. This requirement can be easily enhanced with LLMs, in terms both of hardware and software. In some sense, an horizontal view of hardware scalability could be intended as the addition of an higher number of machines (clusters in a data centre, for example) to manage the workload; conceptually similar, adding more functionalities software-side like adding more libraries or algorithms, can lead to a system able to support more data handling functionalities in the same horizontal view of the business, also activating a more comprehensive LLM-based Agent or creating more than one working together on the same task, or on a different task and converging to determined objectives. Scalability can also be seen vertically, looking at this as a strengthening of infrastructure in the same place (like adding more CPU power to a computer) or making the LLM specialized to a determined task, in the sense of competences and capabilities of the model or in the pure sense of technical data handling, imaging a system that can go deeper in the steps involved in data retrieving, and deeper in the sources the data come from (from example, the model can retrieve the data directly from an IOT device instead of an xls file treated properly before, bypassing the xls creation which would require one or more passage although this situation falls back in the requisite of integration).

Talking about the integration what must be considered is the final objective of the system, as an embedded entity containing different sub-systems more or less complex, that can communicate within and among others. This requisite should want to reach a cohesive workflow, productivity, and operational efficiency, and as the scalability, there are some ways to intend the integration. A common integration reading is vertically, in which the aim is to build a macrosystem able to integrate several functional units into one unit following the silos scheme. A valid example, which is appropriate to explain vertical integration also is the previous one of the specialized LLM – that situation, depending on the point which is chosen to analyse, can be read in both reading keys. Besides vertical integration there is horizontal integration of a system, which is perfectly corresponding to the *Enterprise Service Bus (ESB)*[39] (Yasar K., n.d.) adoption, a subsystem intended to communicate parallelly with other subsystems, embedding all these together.

Another common integration path is the Point-to-Point integration[40], which represents a more direct way to link information systems together, allowing them to directly communicate without an intermediary as for the horizontal integration. This kind of integration is often made using customized API, standards, or direct communication protocols. Apparently easy, this integration can become complex whether the number of nodes is higher since every new connection requires the development of a new set of protocols or interfaces. For an LLM-based Informative System, this way of integration can be deemed appropriate wherever the number of nodes is not too high, since the development of this kind of system perfectly fits these implementations.

---

[39] More information available on: https://www.ibm.com/it-it/topics/esb.
[40] Yasar K., 2022, "What is System Integration? Definition, Methods, Challenges", TechTarger, https://www.techtarget.com/searchcustomerexperience/definition/integration of December 2022.

- Data Safety

Data safety and governance represent fundamental aspects while an information system is engineered, especially whether based on LLMs, considering their capabilities of handling, archiving, and transferring a very large quantity of data and sensible information. Since, unlike a normal information system, which, at least in its simplest and most traditional forms, is isolated from the rest of the world in the company and tampering is only possible by experts who are familiar with the reality, when considering data security in an LLM-based information system, it is necessary to think about the security of the entire system at every level both inside and outside the company boundaries, since the activation and transmission of the data on which to train the model travels over the network to the providers' data centres. The security measures must include physical security, the security of transmission in the network, the application of appropriate encryption or even cryptography measures. Starting with physical security, of utmost importance is the protection of the data centres where the LLM algorithms and associated datasets are hosted, including physical access control to prevent unauthorised access to the server racks, and this control work is much more important in the case of the LLM model provider, as a data breach would involve millions of business operators or individuals. Turning to the network level, security measures must be provided for before, during, and after the transmission of data from the company to the service provider; intrusion prevention measures at network nodes, protection against cyber-attacks of all kinds, or the implementation of a firewall system are just some of the components to be provided for both in the information system in the company and in the modules that are activated. Similarly, the service provider that receives and processes the data must pay the utmost attention to ensure that data breaches do not occur, providing, if necessary, the use of advanced encryption protocols to protect and store the data transmitted by the system, which are applied both to data "at rest" and to data in transit, ensuring that only authorized users can decipher and access the information.

A significantly risky aspect deriving from the implementation of an information system based on LLM is therefore deriving from the transmission of data to the outside, since this flow could open weaknesses in the system and result in data theft or destruction. It will therefore be the task of all the actors involved to create a perfectly safe environment that maximizes the level of protection from external cyber-attacks.

3.3) Costs of developing, usage, and maintenance

One of the factors that makes Generative Artificial Intelligence based on LLMs a technology so capable of being disruptive is the low cost of access and use. When analysing the total cost of such a system, it would be useful to provide a breakdown of the economic effort according to the relevant component, corresponding to two parts, namely System Architecture and Development, which includes the costs of software development and its maintenance, and the cost of using artificial intelligence in the strict sense. On the software development side, creating such a system would entail costs related to the development of the front-end, i.e. the interface the user uses to communicate with the system, back-end, comprising the information architecture behind this system. The development of a web-app like the one proposed here is actually not onerous, considering the countless possibilities of development also thanks to programming languages such as the already presented Python together with its countless free libraries, or thanks to the low-code platforms that are becoming increasingly popular. Thus, the most substantial cost stems from the employment of a developer who has to deal with the creation of the architecture and the consequent engineering side, this always depending on the features required by this system. A simpler system such as the one proposed would not even need this step, especially if used by a few players in the business. What really remains to endure in terms of expenditure is the use of the LLM model held by the provider, which is dislocated with respect to business boundaries. Using such a system makes it possible, as previously mentioned, to bypass the costs related to the investment in hardware and its maintenance, making its procurement easier. What contributes to making these systems truly scalable, in resources and therefore in prices, is their way of functioning based on so-called *tokens*[41], corresponding to the breakdown of a text into smaller parts. Each token, corresponding to a small part of a period, such as a word, is subsequently associated to a number and memorized in a vector, that can be processed and fed in the form of a prompt to the API, which will activate the actual model hosted on the provider's servers. This process is called tokenization and can be done using several algorithms, each of which creates a different dictionary where every token is associated to a specific number. This dictionary is like a map, where, considering how LLMs work, every word are located near the others for similitude; so the word "re" can be associated near "reminiscence", or near "reminder" or even near "react", depending on how this map has been generated and how have these word been mapped and located with.

Given that the LLM model used here is that of OpenAI, some rules are described here, explained in an article written by them, which should help in a simpler interpretation of how the tokens are counted and therefore the costs. To provide a more precise quantitative estimate, as explained by the article, 1 token is equal to approximately 4 characters in the English language, or 75% of a word. Alternatively, 30 tokens are equal to approximately 1-2 sentences, while 2048 tokens are equal to approximately 1500 words. Considering that the dictionary creates different relationships based on the reference language, a sentence does not have the same number of tokens in all languages, as the tokenizer will associate the word with different sentences depending on the language, also changing the number of characters used. Moreover, not only the difference depends on the language used, but all models differ in costs depending on the which is used, so much so that OpenAI's GPT model calculates resources used and costs based on the number of tokens, while Google's chat-Bison model bases the calculation of costs on the number of characters and so on. Below is a price comparison structured by the author of a post on LinkedIn[42] (Lavezzi D., 2023), who compared the models taking

[41] More information available here: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them.
[42] Lavezzi D., 2023, "LLM, ma quanto mi costi?", LinkedIn, https://www.linkedin.com/pulse/llm-ma-quanto-mi-costi-diego-lavezzi-emba-l8zef/?originalSubdomain=it of 12 November 2023.

into account only the costs, depending on the use of the model and not taking further variables into consideration.

Categoria	Modello	Vendor	Prezzo (in $)	
			Input (1k units)	*Output (1k units)*
Text	Jurassic-2 Mid	AI21 Labs	0,0125	
	Jurassic-2 Ultra		0,0188	
	Titan Text – Lite	AWS	0,0003	0,0004
	Titan Text – Express		0,0013	0,0017
	text-Bison	Google	0,001	
	DaVinci	OpenAI	0,02	
	Curie		0,002	
	Babage		0,0005	
	Ada		0,0004	
Chat	Claude Instant	Anthropic	0,0071	
	Claude 2		0,044	
	Titan Text – Lite	AWS	0,0003	0,0004
	Titan Text – Express		0,0013	0,0017
	chat-Bison	Google	0,0005	
	GPT4 (32K context)	OpenAI	0,06	0,12
	GPT4 (8K context)		0,030	0,060
	GPT3.5 (16K context)		0,0030	0,0040
	GPT3.5 (4K context)		0,0015	0,0020
Embeddings	Titan Embeddings	AWS	0,0001	
	textembeddings-Gecko	Google	0,0005	
	AdaV2	OpenAI	0,0001	
	AdaV1		0,004	

Figure 18 - Comparison of available LLM and prices updated to 01/11/2023 (Source: https://www.linkedin.com/pulse/llm-ma-quanto-mi-costi-diego-lavezzi-emba-l8zef/?originalSubdomain=it).

The results of the research that the author is showing in a perfectly summarized manner compare the costs between models of different categories, such as for example the *Text* type models, trained on long texts and particularly effective in the automated writing of contents, the *Chat* models, optimized to be used to automate dialogues and therefore create chatbots for customer service, and the *Embeddings* models, which are particularly useful in text analysis tasks such as sentiment analysis by creating representative vectors. Also noticeable is the cost difference between input and output processing, which means that, for OpenAI's GPT-4 model as example, writing a 1000 token content will costs 0,12 dollars compared to 0,06 dollars required to make it a summary of a content read instead. More generally, without analysing deeper the potential moving a company can take in terms of cost reduction, an estimate can easily say that for an article of 1000 words (approximately 2 to 3 pages). Summarily, trying to estimate how much an interaction, where is answered to the model a summary of a document of approximately 5 pages (2000 words, in Italian), what has to be considered are the embedding cost, consisting of the cost required to have "trained" the model based on the token the document is made of, the asking prompt cost, which takes approximately 10% of the total cost, and the cost computed on output's tokens as:

$$Total\ Cost = \left(\frac{5000\ Token}{1000} * \$\ input\right) + \left(\frac{500\ Token}{1000} * \$\ output\right)$$

Thus, assuming 5000 input tokens (embedding + context + prompt) and 500 output tokens (summary of the text), total costs are compared, for the previously considered model, as follows.

Model	**Vendor**	**Cost (5 pages documents)**
Claude Instant	Anthropic	$ 0,0391
Claude 2		$ 0,2420
Titan Text – Lite	AWS	$ 0,0017
Titan Text – Express		$ 0,0074
chat-Bison	Google	$ 0,0028
GPT4 (32K context)	OpenAI	$ 0,3600
GPT4 (8K context)		$ 0,1800
GPT 3.5 (16K context)		$ 0,0170
GPT 3.5 (4K context)		$ 0,0085

Table 2 - Total cost for summarizing a 5 pages documents, compared for the previously mentioned model (Source: https://www.linkedin.com/pulse/llm-ma-quanto-mi-costi-diego-lavezzi-emba-l8zef/?originalSubdomain=it).

In conclusion, the choice of the most suitable model depends on countless factors. To calculate the costs of a system that is capable of analysing a structured data file such as an xls, xlsx, or csv, one would have to make an estimate of the tokenization of that data, a process that depends on both the amount of data and the complexity of a prompt. In fact, the process underlying customised training generally involves the transformation of information contained in data into natural language propositional information, to which all the mechanisms seen above apply. In any case, the example of the five-page document is in itself significant and sufficiently explanatory.

3.3) Data Protection and legal related aspects

In the age of the Digital Economy, where the transmission of data and information has become a stronghold, generating new production frontiers, business opportunities, and general benefits for the population, one of the biggest challenges to be faced is the issue of personal data protection, both for individuals and legal entities. In today digital world it’s never been easier to gather and store data and it’s never been easier to let it slip by accident or by not taking right precautions. In this sense, some negligent attitudes might potentially drive to some disaster like data breaches, data-destructive hacker attacks, extorsion or more generally serious business threats. Business-side, in treating and archiving confidential data all businesses should always think about client’s right to confidentiality, as a mechanism for brand leveraging and for covenant of the contract also. In this context, the protection of stakeholders' personal data by companies is indeed a legal obligation, but it is also to all intents and purposes an important driver of value for corporate image, through which a company can demonstrate integrity, confidentiality, and instil trust in customers.

All the information gathered about users or businesses can be categorized into three key types of data, all of which must be protected. These are:

1) *Personal Data*: this includes any data that alone or together with any other data relates to an identified or identifiable natural person, such as name, identification number, location data, online identifiers or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identify of that natural person;

2) *Sensitive Personal Data*: this includes personal data that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union

membership, genetic data, biometric data, data concerning health or sexual orientation or data that may facilitate identity theft or payment fraud (including financial account numbers, credit card details and government identification numbers);

3) *Confidential Data*: all the data and information so sensitive that sharing it outside the safety perimeter (which consists of all the actors allowed to access in the business, such as for engagement team or partner or other functional team). Some examples of highly confidential data include trade secrets, merger and acquisition information, income tax return information, sensitive personnel or customer information;

Global institutions are facing the problem of managing personal data with a growing urgency, particularly pushed from the arising of Artificial Intelligence. Collect, gather, handle and transfer data are all operations within the domain of the Digital Economy, such that an absence of regulations can be harmful both for citizens because they would be damaged in their personal sphere, and for businesses because a new growing scenario will be stopped. Internationally, different organizations have started to create and adopt new guidelines regarding personal data protection, although somewhere more heavily than other places; these variations reflect different cultural, economic and political approaches and the following map shows the likelihood of data and privacy regulations, impacting how actors manage data, based on where are they based in the world.

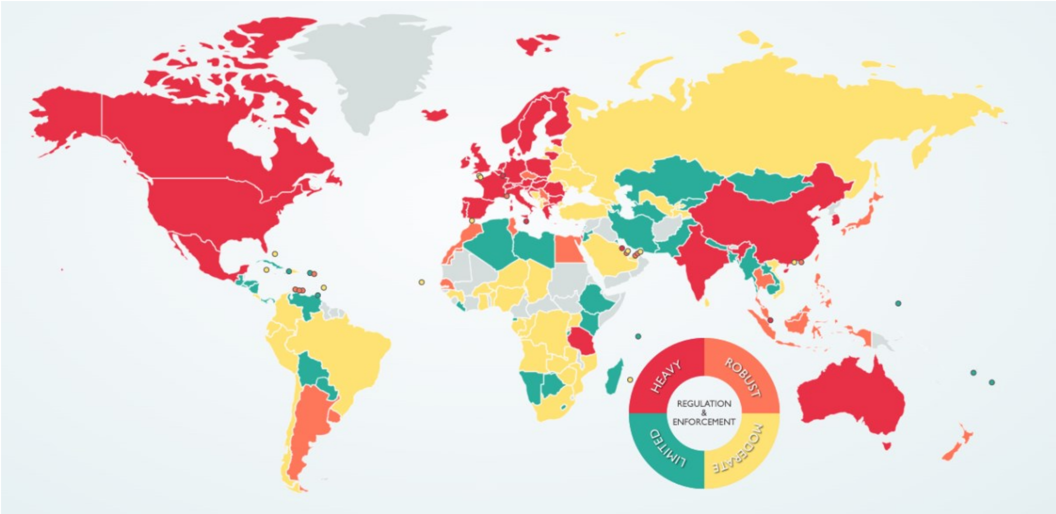


Figure 19 - Different degrees of Data and Privacy Regolations in the world (Source: https://www.dlapiperdataprotection.com/).

Since analysing country by country would be useless here, it is here sufficient to provide a panoramic for countries grouped by continent, distinguishing European Union, Middle East, East (Asia), USA and the remaining Americas. Starting from Europe, this is pioneer in the regulation of personal data establishing elevate standards with Regulation 679/2016[43] (General Data Protection Regulation or GDPR); this regulation fixed the user-centric approach of treating personal data of European countries, placing citizen's protection on the first place. GDPR applies on all the organizations which treat data of UE citizens, independently on where are they located, imposing a strict requirement in terms of agreement, transparency, right to oblivion, and data transferring within and outward European boundaries. This regulation sets out seven principles for the lawful processing of personal data. Processing formally includes the collection, organization, structuring, storage, alteration, consultation, use, communication, combination, restriction, erasure or destruction of personal data, and all these operations must agree with the following principles:

[43] General Data Protection Regulation available here: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=celex%3A32016R0679.

1) Lawfulness, Fairness, and Transparency: any actors involved as controller or responsible of data treatment should be open and honest about who he is and how and why he uses data and make sure any information provided about data processing is accessible and easy to understand;
2) Purpose limitation: any collected data should only be for specified, clear and genuine purposes and none of it should be used for any purpose conflicting with its original. Purposes have to be relevant and reduced only to what is strictly necessary;
3) Data minimization: the data used in the treatment should be reduced to the minimum quantity needed to fulfill the purpose for which they are gathered;
4) Accuracy: as the right of the individuals to get their data treated as right and accurate. Any inaccurate data should be removed;
5) Storage limitation: the controller should not keep the data longer than needed;
6) Integrity and confidentiality (Security): any data processed needs to be treated with care and using the appropriate security measures;
7) Accountability: the responsibilities falls straight to the controller, which has to demonstrate the compliance through appropriate records and measures.

These principles[44] contained in the Regulation frame particular figures involved in data processing, the definitions of which are in Article 4 of the Regulation and which are the *controller*, i.e. the natural or legal person, public authority, service or other body that processes personal data on behalf of the controller; the *processor*, corresponding to the natural or legal person, public administration and any other body, association or organisation responsible for making the basic choices as to the purposes and means of data processing, including the security component the recipient, corresponding to the natural or legal person, public authority, service or other body receiving communication of personal data, whether a third party or not;

[44] An example of principles explained and applied by an institution here: https://www.uhi.ac.uk/en/about-uhi/governance/policies-and-regulations/data-protection/the-seven-principles/.

the *third party*, i.e. the natural or legal person, public authority, service or other body other than the data subject, the controller, the processor and the persons authorised to process personal data under the direct authority of the controller or processor; finally, the central figure on whom the regulation is certainly 'positively biased' corresponding to the *interested* on the treatment, i.e. the natural person where personal data refer to. Also important is the introduction of the concepts of "privacy by design" and "privacy by default" within the regulation, establishing best practices both in terms of construction of the forms containing the request for consent of the interested parties, which must be explicit, sufficiently clear and comprehensible, and provided through unequivocal positive action, both in terms of construction of the platforms with which the user interacts in exercising rights and duties deriving from the relationship with the data controller, declining the application of privacy in the *back-office*, through the adoption of the above principles (for example building a system that keeps user data for the shortest possible time and perhaps automatically deletes them once their function is exhausted) and in the *front-office*, precisely by proposing a form of consent in accordance with the law and set it negative by default, or using standardized icons to explain the focal points of the relationship to the data subject.

Continuing with the comparison of the approach used by different global players, there is the Middle East, which presents an heterogeneous regulation framework depending on the precise location. Some countries, like United Arab Emirates (UAE) and Qatar have introduced specific law for data protection somewhat similar to GDPR but adapted to the local context. As for GDPR, these laws tend to be concentred on data and privacy protection, imposing controller the request of consent for collection, archive and usage of personal data and providing citizens significant right on the control of these data. Regards Asia, however, several various and mixed approaches are adopted, having countries like China or Japan more advanced than others. China, specifically, with its law on data safety and personal information protection, the Personal Information Protection Law of the People's

Republic of China[45] (PIPL), adopts a centralized approach heavily regulated, with an emphasis on data sovereignty, which is detained by the People's Republic of China in a state-centric data approach. On the other hand, Japan with its Act on the Protection of Personal Information[46] (APPI) has released a regulation that aligns more closely with western-style international standards such as the GDPR, adopting a more user-centric approach and generally demonstrating a tendentially opposite orientation to that of China and other countries in the same region, which are gradually developing a growing, albeit slow, awareness of the importance of data privacy.

A cutting-edge approach for which it is interesting to offer a brief overview is that adopted in the United States, which adopts a sectoral approach rather than a unified and all-encompassing regulatory framework, reflecting the traditional preference for verticalized regulation for specific sectors of the economics rather than for general regulatory solutions. The United States do not have a single federal privacy law as Europe does with the GDPR, but relies on a series of federal laws that address specific areas of concern such as the Health Insurance Portability and Accountability Act[47] (HIPAA), which protects patient health information by requiring health care providers and insurers to maintain the confidentiality and security of protected health information, the Children's Online Privacy Protection Act[48] (COPPA), which aims to protect the privacy of children on the web by imposing specific requirements on managers of websites and online services aimed at children under the age of 13 or that collect information relating to them, or the Fair Credit Reporting Act[49] (FCRA), which regulates the collection, distribution and use of consumer credit information, ensuring that the information is treated

[45] Personal Information Protection Law of the People's Republic of China (PIPL) available here: https://personalinformationprotectionlaw.com/.
[46] Act on the Protection of Personal Information available here: https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en.
[47] More information available here: https://www.hhs.gov/hipaa/index.html.
[48] Children's Online Privacy Protection Act (COPPA) available here: https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312.
[49] Fair Credit Reporting Act available here: https://www.ftc.gov/system/files/ftc_gov/pdf/fcra-may2023-508.pdf.

fairly, accurately and confidentially. Alongside these regulations, some states have taken the initiative to develop their own data privacy laws, showing an opposite approach to the unifying one adopted in the European Union. In fact, there has been a growing debate followed by various legislative proposals with the aim of creating a federal law on the protection of user privacy that cuts across all states and is homogeneous throughout the national territory; For this reason, the United States' fragmented approach has raised concerns about its effectiveness in protecting consumer rights in a digital age increasingly dominated by large technology platforms that collect and process massive amounts of personal data, with many criticizing advocates that without consistent federal law, US consumers remain vulnerable to invasive and abusive data collection and processing practices.

In this regard, California has positioned itself as a leader with the introduction of the California Consumer Privacy Act (CCPA) and subsequently with the California Privacy Rights Act (CPRA), which seems to be the only state to have proposed a regulatory advancement truly capable of protect and further expand consumers' rights and businesses' responsibilities regarding the collection, use and sharing of personal data. These represent two important milestones in data privacy legislation in the United States, because they introduce consumer privacy rights and obligations for businesses that are among the most stringent in the country. Starting with the CCPA, this was passed in 2018 and went into effect on January 1, 2020, marking a significant step forward in data protection by giving California residents unprecedented rights to access, control and protect their personal information, such as the right to know, where consumers can ask companies what personal data has been collected, used, shared, or sold, the right to erasure, where consumers can ask companies to delete their personal data , the right to opt-out, with which consumers can oppose the sale of personal data concerning them to third parties or the right to non-discrimination, thanks to which companies cannot discriminate against consumers who exercise the aforementioned rights. However, some critical issues may arise from this regulation deriving from the fact that this does not apply to all companies but to profit-making organizations that have a turnover exceeding 25 million dollars, that buy or sell information of 50,000 or more consumers, families or devices, or those that derive at least 50% of their annual revenue from the sale of consumer personal data. Only with the CPRA, approved in November 2020 and precisely called "CCPA 2.0", has a further strengthening of the protections provided for in the previous CCPA been achieved, with most of the provisions coming into force on January 1, 2023. This regulation not only actually expands consumer rights guaranteed by the CCPA, but introduces new provisions including the establishment of the California Privacy Protection Authority (CCPA) as a government agency dedicated to regulating and enforcing data privacy laws, greater protection for sensitive information, greater emphasis on risk assessment and data security, or the right to correction, where consumers can request correction of their personal data that is inaccurate and was not previously available. The CPRA therefore strengthens California's position as a leader in personal data protection in the United States, introducing standards similar to those of the European GDPR and pushing for greater protection of consumer rights and more responsible management of personal data by companies.

Finally, it is good to also briefly recall that data protection has also been the subject of growing interest in Latin American countries, with countries such as Brazil adopting the General Personal Data Act[50] (Lei General de Proteção de Dados - LGPD), strongly inspired by the GDPR, as well as other countries, which are nonetheless working to strengthen their data protection laws, albeit to a lesser extent, showing a trend towards higher standards of privacy protection. Further insights, although of interest, will not be analysed in this paper.

Always redundant in the topic of protecting the privacy and personal data of users is the issue of the regulation of Artificial Intelligence, currently not yet unified but nevertheless in the process of major evolutionary developments. At the global level, there is still no unified regulation on Artificial Intelligence, with the exception of organisations such as the OECD[51], which has developed the AI Principles (2019)

[50] Brazilian General Personal Data Act (LGPD) available here: https://lgpd-brazil.info/.
[51] More information available here: https://www.oecd.org/digital/artificial-intelligence/.

as a non-binding framework for the responsible development and use of AI, focusing on aspects such as transparency, non-discrimination and security; the G7 has also set up a working group of experts on AI to provide advice and recommendations on AI-related issues, along with various documents reaffirming the leaders' commitment to promoting responsible AI. Across the global landscape, the European Union is the only one to be at the forefront thanks to a series of initiatives aimed both at laying the foundations for an ethical and responsible use of AI and aimed at outlining the vision of the EU as a global leader in investment regulation. In this context, the White Paper on Artificial Intelligence of the European Commission of February[52] (2020) is of considerable importance, which aims to lay the foundations for the use of AI based on the pillars of excellence, investing in research and development by attracting the best talents, of trust, by developing a regulatory framework that ensures the responsible use of AI by protecting citizens' fundamental rights and freedoms, and cooperation, by working with international partners to address global AI challenges. Among the concrete actions identified by the European Commission to support the achievement of these objectives are investments in research and development, the development of a regulatory framework, promoting ethics and responsibility in AI and collaboration with international partners. Specifically for this last point, the EU-US AI Dialogue project was established, a forum to discuss AI issues and to promote cooperation between the EU and the United States. This dialogue is a bilateral forum that aims to promote cooperation on AI research and development, facilitate the sharing of information and best practices, discuss the ethical, legal and social implications of AI, and ultimately coordinate AI policies and regulations. The dialogues are held by leading AI experts and the areas of cooperation range from the safety and reliability of AI to ethics, the impact on the labour market or even the use of AI in specific sectors such as health and environment and safety.

[52] White Paper on Artificial Intelligence of European commission available here: https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_it?filename=commission-white-paper-artificial-intelligence-feb2020_it.pdf.

Following the White Paper, the official proposal for a regulation on Artificial Intelligence arrives in 2021, which aims to create a legal framework for AI in the European Union. The regulation is based on a risk approach, which classifies AI systems according to their level of risk, and the resulting resolutions. High-risk systems will be the first to have to meet a series of rigorous requirements, especially for use in sensitive sectors such as healthcare, justice and law enforcement, and will be subjected to strict controls. The lower risk systems are different and will be subject to less rigorous controls as they are considered safer to use. Following this proposal, the European Parliament and the Council reached a provisional agreement on the Regulation in December 2023, the final text is under negotiation and is expected to enter into force in 2024.

Ultimately, the European Union and the United States have divergent views on personal data protection and user privacy. On the one hand, Europe has taken a rigorous approach with the GDPR, which places the individual at the center and gives them significant control over their data. On the other hand, the United States favors a more flexible system, based on the balance between individual and collective interests, with a predominant role for companies.

Even today, a single solution that satisfies both systems has not been found. Europe is trying to reconcile the protection of privacy with the development of Artificial Intelligence, through specific regulations that aim for "ethical" and responsible AI. In the United States, however, the approach is more pragmatic and market-oriented, with an emphasis on innovation and economic competitiveness. The different settings are reflected in the concrete applications of AI. In Europe, there is a tendency to favour "explainable" and privacy-friendly systems, while in the USA more complex and opaque algorithms are developed, aimed at optimizing results. The challenge for the future is to find a balance between the two visions, guaranteeing both the protection of personal data and the development of responsible AI that is beneficial to society. An open and constructive dialogue between Europe and the United States will therefore be fundamental to achieving this goal.

Chapter 4) Conclusion and Future directions

This thesis explored the application of Generative Artificial Intelligence in the business context, with a particular focus on increasing the efficiency of operational processes and improving the quality of decision-making in management and governance. Through the development and testing of an advanced system, implemented in Python by a non-programmer user, this research demonstrated that it is possible to facilitate the direct interaction of business users with data, by means of an intuitive interface that allows them to formulate complex queries and receive answers in the form of numerical analyses and interactive graphical visualisations.

The methodology adopted involved the use of advanced programming techniques for natural language processing and data visualisation, allowing non-expert users to interact with the system through text prompts in a completely new way. This made it possible to assess the effectiveness of a low-code approach in breaking down technical barriers for non-programming users, while maintaining deep integration with the specific information needs of the business context. This approach has seen the implementation of a medium-complexity system by a user not purely verticalized towards software engineering, demonstrating the concrete possibility of developing low to medium complexity software by a business-side user. The need for new professional figures that are transversal between management and the new technologies has thus been demonstrated here, with a practical example of how these can be capable of generating considerable added value in business activities, thus enabling increased communication between various company areas and various skills, in today's ever-changing context.

As a result, this work proposes a replicable model for the development of artificial intelligence-based applications or information systems that are flexible, highly customisable, and integrable, and ultimately scalable both technically and in functionality. Such a system, with its innumerable potential uses, is proposed as capable of generating added value as never before, by means of new, highly intelligent virtual assistants that can be of extreme help in the most diverse business contexts, as demonstrated in the examples. Many companies are being proactive during this phase of epochal change, investing considerable resources in research, development, and training of individuals, testing the use of these systems directly in daily activities, and experiencing increased levels of productivity. The ability of these systems to understand natural language and translate it in noticeable and high creative results will certainly be appreciated on a large scale, as has already happened for other General Purpose Technologies, such as the Internet, which changed irreversibly our lives.

In conclusion, the evolution of Generative AI is radically transforming the work and knowledge landscape, underlining a paradigm shift that should not be underestimated. This technology does not merely simplify existing processes through automation but extends creative and analytical possibilities to a much wider audience, making knowledge and technical skills more accessible. Generative AI tools are democratising the ability to create, analyse and learn, removing previously insurmountable barriers and enabling more and more people to contribute meaningfully to different work and creative domains. This democratisation not only increases individual empowerment but also stimulates innovation and growth in areas that were previously dominated exclusively by highly specialised experts. The ability to turn ideas into reality, regardless of a person's technical or creative background, is a driving force for a future where innovation is more inclusive and distributed. Consequently, the adoption of generative AI paves the way for a world in which the value of work and knowledge evolves, moving increasingly towards the integration of humanity and technology, and where human potential is amplified by artificial intelligence in previously unimaginable ways. The impact of this technology on the world of work and learning is a clear indicator of how the future can be shaped more equitably, efficiently, and creatively through the empowerment provided to individuals of all skill levels.

References

Ana Aizcorbe, S. K. (2004). Moore's Law and the Semiconductor Industry: A Vintage Model. *Moore's Law and the Semiconductor Industry: A Vintage Model*.

AWS Amazon. *Cos'è Cloud Native?* Retrieved from AWS Amazon: https://aws.amazon.com/it/what-is/cloud-native/#:~:text=The%20cloud%2Dnative%20architecture%20combines,blocks%20of%20cloud%2Dnative%20architecture

Brynjolfsson E., L. D. (2023). Generative AI at work. *NBER Working Paper Series*. Retrieved from https://www2.nber.org/system/files/working_papers/w31161/w31161.pdf

Child R., G. S. (2019, April 23). *Generating Long Sequences with Sparse Transformers.* Retrieved from Cornell University: https://arxiv.org/abs/1904.10509v1

DatabaseCamp. (2021, December 14). *What are Recurrent Neural Networks?* Retrieved from Database Camp: https://databasecamp.de/en/ml/recurrent-neural-network

Fan X., T. P. (2022, February 11). *An Introduction to Convultional Neural Network (CNN)*. Retrieved from Medium: https://medium.com/sfu-cspmp/an-introduction-to-convolutional-neural-network-cnn-207cdb53db97

Feng-hsiung H. (1999). IBM's Deep Blue Chess grandmaster chips. *IEEE Micro*, 70-81.

Ferrario C. (2022, November 20). *Reti neurali artificiali: cosa sono, come funzionano e perché vengono costruite*. Retrieved from Geopop: https://www.geopop.it/reti-neurali-artificiali-cosa-sono-come-funzionano-e-perche-vengono-costruite/ of 20 november 2022

GitHub. *GitHub*. Retrieved from https://media.githubusercontent.com/media/datablist/sample-csv-files/main/files/organizations/organizations-10000.csv

Google DeepMind. *Gemini*. Retrieved from Gemini: https://deepmind.google/technologies/gemini/

International Business Machines Corporation. (n.d.). *Neural Networks*. Retrieved from https://www.ibm.com/it-it/topics/neural-networks.

Jermakowicz E. (2023). The Coming Transformative Impact of Large Language Models and Artificial Intelligence on Global Business and Education. *Journal of Global Awareness, St. John's University*.

Juhás G. et al. (2022). Low code platforms and languages: the future of software development. (pp. 286-293). IEEE Conference - 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA).

Keim R. (2019, November 24). *How to Train a Basic Perceptron*. Retrieved from All About Circuits: https://www.allaboutcircuits.com/technical-articles/how-to-train-a-basic-perceptron-neural-network/

Lavecchia V. *Sistemi centralizzati e sistemi distribuiti*. Retrieved from Informatica e Ingegneria Online: https://vitolavecchia.altervista.org/sistemi-centralizzati-sistemi-distribuiti/

Lavezzi D. (2023, November 2023). *LLM, ma quanto mi costi?* Retrieved from Linkedin: https://www.linkedin.com/pulse/llm-ma-quanto-mi-costi-diego-lavezzi-emba-l8zef/?originalSubdomain=it

Luo Q., Puett M., & Smith M. (2023). A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and Youtube. *A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and Youtube*.

Riccò B. (2008). *Legge di Moore*. Retrieved from Treccani: http://www.treccani.it/enciclopedia/legge-di-moore_%28Enciclopedia_della_Scienza_e_della_Tecnica%29/

Son H. (2023, September 18). *Morgan Stanley kicks off generative AI era on Wall Street with assistant for financial advisors*. Retrieved from CNBC: https://www.cnbc.com/2023/09/18/morgan-stanley-chatgpt-financial-advisors.html

Tarver E. (2021, September 01). *What Are the Primary Activities of Michael Porter's Value Chain?* Retrieved from Investopedia: https://www.investopedia.com/ask/answers/050115/what-are-primary-activities-michael-porters-value-chain.asp

Tosato M. (2013, April 30). *Addestramento reti neurali feed-forward multi-layered tramite Error Backpropagation*. Retrieved from Paperblog: https://it.paperblog.com/addestramento-reti-neurali-feed-forward-multi-layered-tramite-error-backpropagation-1988250

Tseng P.K. (2023, March 01). *TrendForse Says with Cloud Companies Initiating AI Arms Race, GPU Demand from ChatGPT Could Reach 30,000 Chips as It Readies for Commercialization*. Retrieved from TrendForce: https://www.trendforce.com/presscenter/news/20230301-11584.html

Vaswani A. et al. (2017). Attention Is All You Need. *Attention Is All You Need*.

Vincent P. et al. (2021, April 13). *Identify and Evaluate Your Next Low-Code Development Technologies.* Retrieved from Gartner: https://xone.es/wp-content/uploads/Identify_and_Evaluat_742070_ndx.pdf

Wikipedia. *Information Theory*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Information_theory

Wulf J., M. J. *Towards a Taxonomy of Large Language Model based Business model Transformation*. Zurich.

Yasar K. *What is System Integration? Definition, Methods, Challenges*. Retrieved from Techtarget: https://www.techtarget.com/searchcustomerexperience/definition/integration