



UNIVERSITÀ DEGLI STUDI DI BERGAMO

Dipartimento di Lingue, Letterature e Culture Straniere

Corso di Laurea Magistrale in Planning and Management of Tourism Systems

Classe n. 49 - Progettazione e Gestione dei Sistemi Turistici

Advancing tourism research with Artificial Intelligence: integrating Large Language Models for searching and extracting textual data from PNRR documentation

Relatore:

Chiar.mo Prof. Nicola Cortesi

Correlatore:

Chiar.mo Prof. Ángel Herrero Cresp

Tesi di Laurea Magistrale

Fatemeh KAZEMIANROUHI

Matricola n.

1087052

*As our case is new,
we have to think and act in a new way*

– Abraham Lincoln

Acknowledgements

Writing the acknowledgments for this thesis has been more challenging than writing the thesis itself, as they represent a heartfelt reflection of the people who made this journey possible. This work would not have been possible without the incredible support, guidance, and collaboration of so many individuals who stood by me. I owe them my deepest gratitude, for without their involvement and encouragement, I would not have been able to bring this thesis to life. As a student who has faced the complexities of both academic challenges and the personal struggles of being a migrant, I am profoundly aware that this achievement is not mine alone, but the result of a collective effort.

As the formal representative of PMTS students, I dedicate this work to all my **fellow students** and **migrants** who, like me, have had to balance their academic ambitions with the practical realities of life in a new place. The challenges we face are numerous - navigating cultural barriers, adapting to new systems, and often balancing our studies with work. Yet, it is our shared resilience and determination that has allowed us to push through these obstacles. This acknowledgment is a tribute to every student who has struggled, persevered, and succeeded in this journey, proving that with enough determination and support, even the hardest of paths can be walked.

Finally, I want to express my deepest gratitude to those closest to me - my **family**, who provided me with unwavering emotional and moral support; my **professors**, who offered their knowledge and mentorship, and most especially my **supervisor** and co-supervisor, whose patience, guidance, and encouragement steered me in the right direction when I needed it the most. To everyone who contributed to my studies and to where I stand today, this acknowledgment is a humble tribute. Your belief in me, your insights, and your encouragement have been invaluable. I am eternally grateful to each one of you for shaping my journey and allowing me to reach this significant milestone.

Table of contents

Introduction	6
1. Large Language Models (LLMs)	10
1.1 Brief history of Large Language Models	11
1.2 How Large Language Models work	13
1.2.1 From biological neurons to artificial neural networks	14
1.2.2 The Transformer architecture	19
1.2.3 Model training and output	25
1.3 Main limitations of LLMs	26
1.4 LLMs and human consciousness	28
1.5 Evolution of GPT models	30
1.6 Other notable LLMs for analyzing PDF	33
1.7 Common applications of LLMs	36
1.8 Principles of prompt engineering	39
1.9 Integrating ChatGPT in scientific research	41
2. Main applications of AI and LLMs to the tourism sector	44
2.1 Personalization of the touristic experience	46
2.2 Digital marketing strategies in tourism	48
2.3 Market research in tourism	51
2.4 Virtual reality experiences in tourism	53
2.5 Online reputation management in tourism	54
2.6 Analysis of customer feedbacks in tourism	55
2.7 Crisis management in tourism	57
2.8 Price optimization in tourism	59

3. Case study: PNRR and regeneration of Italian towns	62
3.1 The ‘Town Attractiveness’ call	63
3.2 Projects of Line B of the Town Attractiveness call	66
3.3 Project proposals	70
4. Methodology	74
4.1 Data extraction	74
4.2 OCR conversion	81
4.3 List of municipalities	84
5. Results	88
5.1 Extraction of town interventions	88
5.2 Validation of administrative information	90
5.3 Error sources of ChatGPT-4	92
Conclusions	96
Appendix	100
Acronyms	106
Bibliography	108

Introduction

In an era where data is the new currency, the tourism sector now stands at a crossroad, a point of convergence between Artificial Intelligence (AI) and Big Data (Green, 2022). Information technologies allow to forecast many variables of interest, but they also introduce so many revolutionary advances that, paradoxically, make it impossible to forecast the evolution of society itself, or just that of the tourism industry.

Generative AI models known as Large Language Models (LLMs) were recently introduced to process large quantities of textual information and are a new and powerful tool at disposal of the scientific community (Evans & Patel, 2020). These advanced models, with their robust natural language processing capabilities, represent a new paradigm shift (Li & Zhang, 2021).

A common bottleneck in tourism research is the laborious task of searching through massive volumes of touristic data to extract relevant insights and compile them into cohesive datasets. This foundational step remains a manual, resource-intensive endeavor, monopolizing significant human effort and detracting from time available for analyzing results (Smith & Johnson, 2020). Nowadays, computer scientists are required to support many fields of research, but most of their tasks may be soon automated by LLMs.

The tourism industry encompasses different segments like travel, hospitality, cultural heritage, and leisure, and it is inherently dynamic and data-rich (Doe et al., 2019). It continually spawns an immense array of data, including textual, numerical, visual, and audio content, through various channels like online reviews, customer feedback, social media, travel blogs, and forum discussions (Miller & Davis, 2018). This data is crucial for strategic decision-making in tourism. Yet, traditional data processing and analysis methods often fall short in tackling the sheer volume and complexity of this data, leading to untapped potential (Taylor & White, 2017).

A key feature of LLMs is the ability to serve as virtual assistants or ‘digital

programmers', converting vocal or written commands into programming languages like Python or Java, and providing detailed help in understanding the inner working of the command themselves (Khan & Ali, 2019). This capability now empowers students and researchers alike with a degree computational proficiency akin to that of a seasoned computer scientist (Jones, 2023).

This thesis aims to understand and examine the profound impact of LLMs on data mining and analysis in the tourism sector, a domain abundant with unstructured textual and graphical data, traditionally more challenging to analyze than numerical data (Wang & Chung, 2022). It explores the usefulness of a particular class of LLMs, called Generative Pre-trained Transformers (GPT), in improving data mining and analysis of textual information (Peterson, 2021).

The present study is divided in two parts. The first one is dedicated at establishing a foundational understanding of LLMs, tracing their evolution, underlying working, and an in-depth examination of their applications, particularly for tourism services. Chapters 1 and 2 focus on these topics. Following this overview, the second part of the thesis (chapters 3, 4 and 5) delves into a concrete application of GPT to the tourism industry: the automatic extraction and generation of a touristic database. The goal of the second part is the creation of new dataset, obtained by searching and retrieving useful information stored in hundreds of very long textual documents published in the framework of one of the most important national tourism tenders of the last decades: Town Attractiveness tender ('Attrattività dei Borghi'). This one-billion euro tender was aimed at the regeneration of rural areas and towns below 5000 inhabitants. It was funded by the National Recovery and Resilience Plan (PNRR) and assessed by the Italian Ministry of Culture (Nature, 2022).

PNRR is Italy's strategic response to the global COVID-19 emergency, addressing the challenges that have hindered the country's economic, social, and environmental development over the year. It forms a crucial part of the European Union's recovery mechanism, Next Generation EU, which is characterized by its significant scope and ambition, allocating substantial resources for recovery efforts.

Thousands of Italian towns engaged in the PNRR tender 'Attrattività dei Borghi', submitting proposals to rejuvenate and repopulate their regions. The volume of data

provided was immense, prompting the Ministry of Culture (MiC) to make it publicly available, an infrequent but desired property of MiC's public tenders. This information is encapsulated in hundreds of extensive and intricate PDF documents, one for each of the 221 towns promoted by the MiC for the PNRR funding, making it challenging to transform and structure this data into a readily usable dataset format.

In this thesis, GPT-4, the most advanced iteration of the GPT as of the writing date, was employed to extract valuable tourism-related information from those intricate PDF documents. This was achieved by relying just on a few skillfully crafted prompts of the GPT interface, the chatbot known as 'ChatGPT'. The prompts enabled the automatic extraction of all data regarding the type of interventions proposed by each town, along with their budgets.

To evaluate ChatGPT's skill in extracting textual data from the documents, the data it generated were benchmarked against those manually compiled for a prior master's thesis by Bonera (2023): 'Sustainable tourism and regeneration of Italy's hamlets: comparative analysis of two PNRR projects in Lombardy'. Bonera's work involved a similar process of analyzing and extracting data from the same PNRR documents, albeit without leveraging Large Language Models (LLMs). This comparison enabled a detailed assessment between the manual and ChatGPT-produced datasets. A notable distinction is that while Bonera's dataset encompassed only the 29 towns selected by the MiC in Lombardy, the complete dataset should include all 221 Italian towns chosen by the tender. This extended dataset should be produced in significantly less time than that required by human efforts, and it is the primary reason for undertaking this thesis.

In summary, this thesis investigates the application of ChatGPT-4 in advancing tourism research by searching, extracting, processing and analyzing data from the PNRR documentation. Its main focus is on extracting and organizing vast amounts of data from the PNRR's Town Attractiveness tender, a significant challenge due to the extensive and complex nature of the information. ChatGPT-4's role in transforming hundreds of heterogeneous PDF documents into a comprehensive dataset is assessed, comparing the results with a manually created dataset from a previous thesis. Such a comparison aims to measure GPT-4's efficiency in handling larger data volumes more quickly than traditional methods, hoping of heralding a new era for data analysis in tourism.

1. Large Language Models

Any LLM just predicts the next word of a sentence, one by one, until the full desired text is generated. The forecast is not deterministic but probabilistic: if the prediction is repeated, there is a certain probability that the forecasted word will be different from the previous prediction (Brown et al, 2020).

At first glance, one may think that the task of predicting the next word of a sentence is much simpler than that of generating a whole page of new text: a story, a poem, or a full answer to a question. Indeed, the first answers generated by the earlier versions of LLMs didn't make a lot of sense. However, once the number of parameters of LLMs grew exponentially, all in a sudden the answers become quite reasonable (Lakshmanan, 2022).

One of the most astonishing features of LLMs is that they exhibit the so-called *emerging skills*: new, unintended properties that manifest by themselves just by increasing the model size, even if LLMs were not explicitly developed to generate whole sentences and answers, but just one word at time (LeCun, Bengio, & Hinton, 2015).

The first chapter of this work is focused on the description of LLMs, illustrating the history of their development (section 1.1), briefly explaining their inner working (1.2) and clarifying their relationship with human consciousness (1.3). Then, the evolution of one particular class of LLMs called GPT (1.4) is shown, followed by the list of the most prominent LLMs able to analyze new documents provided by the users at the moment of writing (1.5).

An overview of the most common applications of LLMs is presented in section 1.6, along with the best practices in order to obtain the most useful results from any LLM (1.7). Chapter 1 ends with a discussion of the principles of Prompt Engineering (1.8) and of the benefits of integrating LLMs for writing scientific articles and publications (1.9).

1.1 Brief history of Large Language Models

The history of LLMs is characterized by continuous innovation and progress, driven by advancements in machine learning, non-linear programming (NLP), and computational resources. The evolution of LLMs reflects the continuous innovation and progress in NLP research. From the early days of rule-based systems, to the advent of modern transformer-based models, LLMs have significantly evolved into powerful tools for processing and understanding human language.

Their journey can be traced back to the early days of computerized conversation agents, notably with the advent of Eliza, a groundbreaking program developed by Joseph Weizenbaum in 1966. Though rudimentary by today's standards, Eliza showcased the potential for machines to engage in natural language conversations. Eliza operated using pattern-matching techniques, simulating a psychotherapist and responding to user inputs in a conversational manner, as shown in figure 1.

Eliza marked a pioneering step in the realm of computer-mediated communication, because it was the first time that a programmer developed a human-machine interaction with the aim of creating the illusion (albeit brief) of a human-human dialogue.

```
Welcome to
                EEEEE LL    IIII  ZZZZZ  AAAAA
                EE    LL    II     ZZ    AA  AA
                EEEEE LL    II     ZZZ   AAAAAA
                EE    LL    II     ZZ    AA  AA
                EEEEE LLLLLL IIII ZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Figure 1: Main screen of Eliza, the first example of a bot that simulates dialogue with humans. Source: https://it.wikipedia.org/wiki/ELIZA_%28chatterbot%29

Subsequent decades witnessed a shift from rule-based approaches to statistical models. Researchers explored diverse methodologies, including Hidden Markov Models and n-gram models. The emergence of neural network architectures, coupled with the availability of vast datasets, revolutionized the field of non-linear programming (NLP). Deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), enabled to learn complex patterns and representations from raw text data (LeCun, Bengio, & Hinton, 2015). These advancements laid the foundation for the development of modern LLMs with enhanced language understanding and generation capabilities.

It was not until the last decade, with the advent of machine learning techniques (particularly deep learning), that LLMs began to demonstrate unprecedented capabilities. The key milestone in the evolution of LLMs was the introduction of the Transformer architecture by Vaswani et al. in 2017. It was based on ‘self-attention’ mechanisms, and achieved state-of-the-art performance in various NLP tasks by capturing long-range dependencies and contextual information in text. The Transformer architecture served as a catalyst for the development of transformer-based LLMs, such as OpenAI's GPT series, which demonstrated remarkable proficiency across a wide range of language tasks and set new benchmarks for natural language understanding and generation (Radford et al., 2018).

The evolution of LLMs, from Eliza to contemporary transformer-based models, showcases the remarkable progress in natural language processing. Since 2018, the number of parameters of LLMs increased exponentially from the 94 millions of ELMo LLM in 2018 to the 530 billions of Megatron-Turing LLM in 2021 (see figure 2). More recently, OpenAI's ChatGPT-4 was trained with more than one trillion parameters (the exact number is undisclosed). A single parameter represent the strength of the connection between two artificial neurons, and can be interpreted as the fundamental unit of memory of the system, as explained in section 1.2.

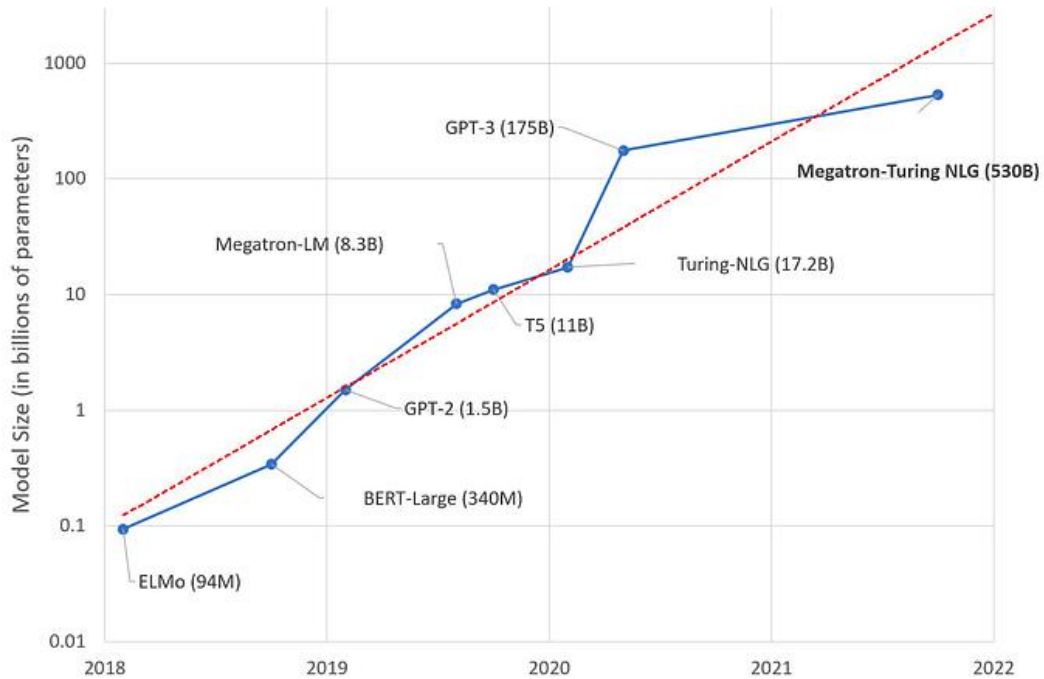


Figure 2: increase of the number of parameters of the main LLMs developed from 2018 to 2022. The number of parameters of each model is indicated both in the vertical axis and in parenthesis. Vertical scale is not linear, but logarithmic (powers of 10). Red line show the long-term exponential trend. Source: *Analyticsvidhya.com*

1.2 How Large Language Models work

LLMs like GPT-4 represent the pinnacle of current AI technology, especially in the field of natural language processing. The sophistication of these models lies in their ability to understand and generate human language with remarkable coherence and relevance (Brown et al, 2020).

The development of LLMs required knowledge in different scientific areas, from neurophysiology to linguistic, from scientific programming to parallel computing. Theoretical understanding of artificial neural networks was already fully established at the end of the last century; however, in order to fully implement them, a few technological advancements were needed, in particular the adoption of Graphic Processing Units (GPU) for scientific computation (Raina et al., 2009). Parallel processing of LLMs with GPUs become available during the last decade, dramatically dropping the time to train LLMs from several million years to a few months. This is the main reason of the long wait to see the first LLMs in action (Dean et al, 2012).

1.2.1 From biological neurons to artificial neural networks

It all started during the second half of the XIX century, with a few key discovery in the field of the neurosciences, in particular about how electrical currents propagates in the brain. The first important discovery was that neurons behave as conductors of small currents of a few mA, much lower than those of LED bulbs (von Helmholtz, 1887). Thus, groups of interconnected neurons can be considered a microscopic electrical circuit.

Figure 3 illustrates the shape of the average human or animal neuron. It is characterized by a central cell body (called *soma*), containing the nucleus and other apparatus like the mitochondrion (present in every human cell). The main difference between neurons and other cells is that their body is surrounded by many branches, called *dendrites*, and by one and only one particularly long branch, called *axon*. Average axon length is 1 cm, which is indeed very long compared to the brain size. Dendrites are similar to tree branches, in the sense that they can easily form really complex structures that resembles those of trees. Each neuron, in fact, can have anywhere from just a few dendrites to several thousands of them (Ramón y Cajal, 1899).

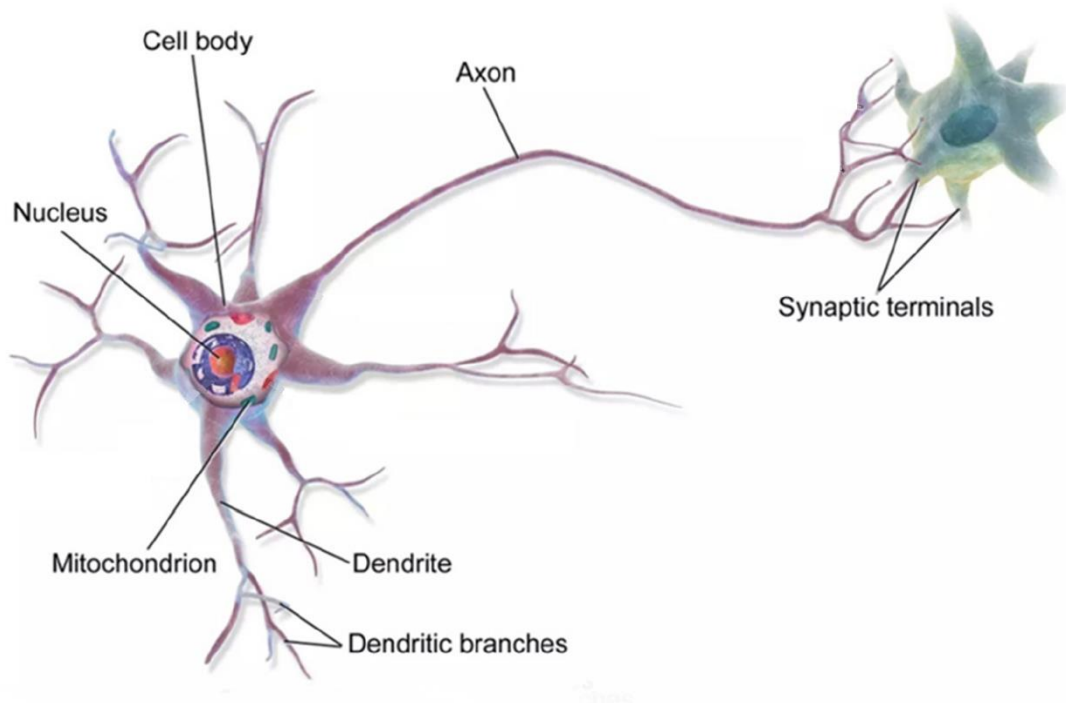


Figure 3. Graphical representation of the average human neuron, with its main body (or soma), its dendritic tree and its long axon that projects to other neurons. Source: *Sui et al. (2020)*.

Most of the dendrites are also reached by the axons of one or more neurons; they do not physically touch between them, but they are close enough to form a tiny structure called *synapse*. Inside the synapse, the electrical signal that crosses its nearby axon is converted into a chemical signal, in the form of hundreds of neurotransmitters that are released from one side of the synapse. Common neurotransmitters include glutamate, GABA, dopamine, serotonin, and acetylcholine (Dale, 1935).

When they reach the other side of the synapse, the neurotransmitters bind themselves to specific receptors on the membrane of the dendrite, causing a change of its electric potential and initiating a new electric current in the dendritic tree of the new neuron, that eventually (after a few nanoseconds) reaches its central body. Synapses are not static structures; they can change over time, in response to increases or decreases in their activity (Hebb, 1949). This plasticity is the cellular basis for learning and memory, as later shown in this section.

The central body of the neuron is crossed by the electrical signals coming from all its dendrites. In case the total signal exceeds a certain thresholds (called ‘action potential threshold’), the neuron fires a new electrical signal that propagates along its axon. The axon connects the neuron to other ones. It can also end with several branches that connect to thousands of neurons (as in the case of the Purkinje cells in the cerebellum), but always through the same structures, the synapses. It's been estimated that the human brain could have around 100 trillion (10^{14}) synapses in total, although this number could vary among individuals (Shepherd, 1998).

Such a propagation is not so difficult to simulate mathematically: the intensity of the electrical current propagating in the axon number 0 can be represented by the number x_0 , while the intensity originated from axon number 1 can be represented by the number x_1 , and so on until each axon has its associate intensity (see Figure 4 below). The synapse modulates the original current coming from the axons, by a factor w (called ‘weight’), so that the current in the dendritic membrane is not exactly the same one of the axon, but it is weighted (multiplied) by the factor w_i , where i indicates the number of the synapse or of the dendrite (McCulloch et al., 1943).

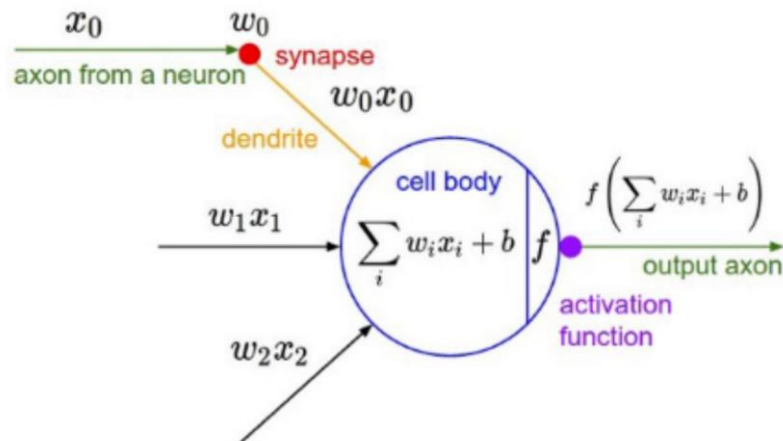


Figure 4. Schematization of the average human neuron shown in the previous figure. The dendritic tree is replaced by the currents x_i , that are weighted by w_i , the factor representing the influence of each synapse. All the electrical currents coming from the dendrites reaches the central body of the neuron, which itself possess the constant electrical potential b , that sums up to the total. The resulting current is modified by the activation function f (e.g: the sigmoid function); the new current propagates along the axon. Source: <https://steemit.com/technology/@davidfumo/a-gentle-introduction-to-neural-networks>

The artificial neuron described in the previous figure is also called *node*. All currents converge in the central part of the node (the soma), which experiences the algebraic sum of all of them, plus a constant term indicated as b that represents the starting value (or *bias*) that does not depend on the influence of other nodes. This concept is a fundamental aspect of the neural network model, as described in Rosenblatt (1958).

Finally, function f is applied to the total current, usually to normalize it to a more adequate range of values (e.g.: the numbers between -1 or 1, or only positive numbers). Such a function is called ‘activation function’. In our brain, there is only one type of activation functions, as the neuron fires if and only if the total current exceeds a certain threshold.

In an artificial neural network, there are many different type of possible activation functions, depending on the specific task required. The most commonly employed activation function is called ‘sigmoid’ function; it converts the input number x into a new one, that is always in the range from 0 and +1, in a non-linear way (see figure 5). This function and its relevance to neural networks are elaborated in Cybenko (1989).

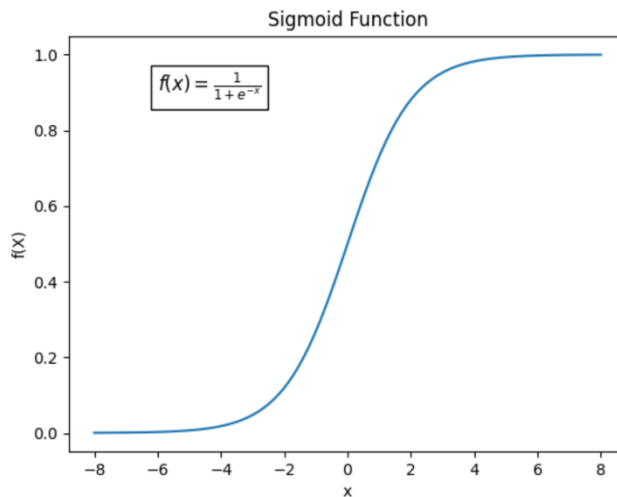


Figure 5. The sigmoid function employed by many LLMs in order to normalize the input values of each artificial neuron to a number from 0 and 1, before sending the signal to the following neurons. Source: <https://www.codecademy.com/resources/docs/ai/neural-networks/sigmoid-activation-function>

There are two main types of artificial neural networks (ANN): simple ANN and Deep Learning ANN, also called Deep Neural Networks or DNN. The difference between them lies in the number of layers: simple ANNs typically have three layers in total, while DNNs have at least six layers (see Figure 6).

A layer is made up of two or more nodes (neurons) that are all connected to the previous layer and to the following one in the network, but not within them. These layers are not mere passive conduits of information; they actively process and interpret the data flowing through them.

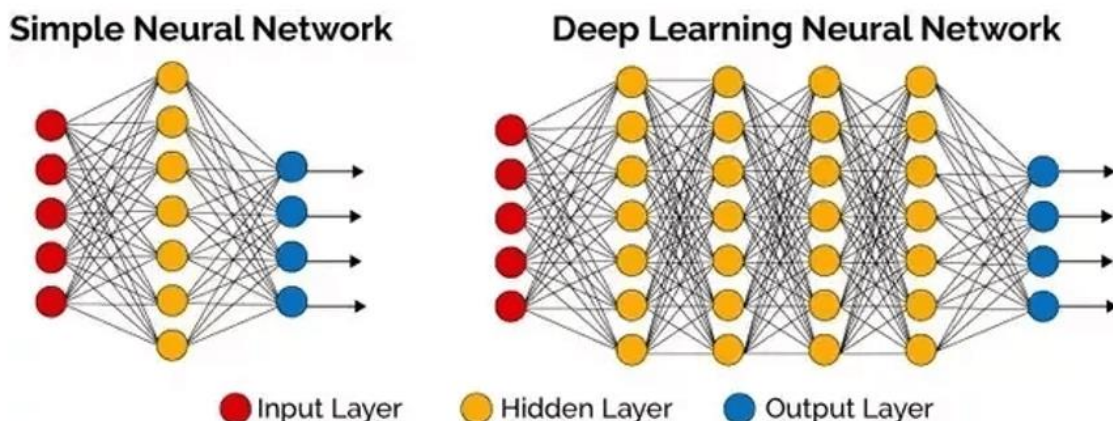


Figure 6. Difference between simple neural networks (left) and Deep Learning ones (right). Source: https://www.researchgate.net/publication/335856901_Survey_on_Intrusion_Detection_Systems_based_on_Deep_Learning

Each node of these layers can be thought of as a mini-processor, working in concert with others to decode the nuances of language. For instance, ChatGPT-3.5 is based on 96 layers, thus it is considered a type of DNN. This distinction and the architecture of DNNs are detailed in Schmidhuber (2015).

In simple ANNs, the limited number of layers restricts the complexity of the tasks they can perform, as explained in Haykin (1994). On the other hand, DNNs, with their increased number of layers, are capable of handling more complex tasks and extracting higher levels of abstraction from data, as discussed in LeCun et al. 2015. This ability makes DNNs suitable for a wide range of applications, including natural language processing as employed in models like ChatGPT-3.5.

The input layer (red points in figure 6) is made up by a type of neurons that do not process data in any way: they are just used to load the data in the ANN and send it to the first hidden layer to process (yellow points). For example, each node could store the information about a single pixel of an input image, or a single word of a text page.

All computations are performed in the so-called ‘hidden layer’, which is made up by a single ‘column’ of node in case of simple ANN, or more than one in case of DNN. Each node of a hidden layer is connected to all nodes of the previous layer and to all nodes of the following layer. The number of nodes per layer depends on the complexity of the task, and can be very high. For example, ChatGPT-3.5 has about 2 billion nodes per layer.

The results of the internal computations are then passed to the last layer of the ANN, that is called the output layer (blue points in figure 6). Depending on the type of ANN, this nodes can represent just a single number, a True/False answer, a sequence of probabilities, a word or an image. It is interesting to notice that the total number T of connections of a DNN is equal to the number of layers L , per the number of nodes per layer N , per the number of connection per node (that is also equal to N). Thus,

$$T = L \times N^2 \tag{1}$$

For example, in case of ChatGPT-3.5, T is equal to 175 billion connections (also called ‘parameters’). ChatGPT-4 should have more than one trillion parameters, although the real number was not publicly available at the time of writing. For a comparison, the number of connections of the human brain (the synapses) is about 100 trillions. On the basis of this number, the average adult human brain might have the ability to store the

equivalent of approximately 2500 terabytes of digital memory, as estimated by Drachman (2005). Note that the biggest hard drive to date can only store 10 terabytes, which pales in comparison to the storage capacity of the human brain.

All the connections in a biological neural network (dendrites and axons) are fixed and cannot change neither move or grow as time passes. Thus, the only way to learn and remember new information is to store it in the only place of the brain that can alter its properties during all life: the synapses. Both learning and memory are processes that require a constant update of information, so our brain can only store this data in the synapses. This concept is explained in detail in Kandel et al. (2013). In an ANN, synapses are represented by the weights w_i . Thus, the weights of an ANN can also be considered not only as simple parameters, but as the true memory of the system (McClelland et al., 1986).

Many examples of working ANNs are publicly available on GitHub, the most employed worldwide repository of scripts. Readers interested in programming their own ANN, can already take advantage of the existing code shared. We recommend starting with a simple ANN devised to recognize numbers, not text, as this task can be easily realized in a few dozen rows of code in Python language (for instance, the script ‘nn.py’ available at <https://github.com/Bot-Academy/NeuralNetworkFromScratch> and clearly illustrated step by step in the following YouTube video published by the Bot Academy: <https://www.youtube.com/watch?v=9RN2Wr8xvro&t=3s>). In this way, readers will realize that the complexity of ANNs is not determined by the efforts required to program the algorithm, but mainly by the sheer computational power required to execute it.

1.2.2 The Transformer architecture

In order to grasp the inner working of LLMs, it's essential to delve into the various stages of their development and operation, from the initial data ingestion to the neural network architecture and the processes of learning and text generation.

The journey of an LLM begins with the ingestion of a massive corpus of freely available text data taken from the World Wide Web. This data is incredibly diverse, spanning a multitude of genres, styles, and subjects. From literary classics to contemporary web articles, and from technical manuals to everyday conversational text, the range is exhaustive. This diversity is critical, as it exposes the model to the myriad

ways language is used across different contexts (Brown et al., 2020).

During the data preparation phase, input text is meticulously processed in order to remove repeated sentences, non-sense text and toxic text (hate speeches, racist expressions or pornographic material). Each piece of text is broken down into smaller units known as *tokens*, which can be words or parts of words, or even syllables or punctuations marks like commas or points. This last method is particularly useful for handling new or technical words and for reducing the vocabulary size. For example, the word "unexpected" might be broken down into "un-", "expect", and "-ed". This ‘tokenization’ process is vital, as it transforms the raw text into a format that the model can efficiently process (Sutskever et al, 2014). GPT models use an approach called Byte Pair Encoding (BPE), effective in handling both a large vocabulary and a manageable number of tokens. In languages that don’t use spaces to separate words (e.g: Chinese), each individual character is considered a token.

Each token is then transformed in an integer number, in order to represent it internally. Such a number usually represent the position of the word inside the vocabulary (figure 7). Thus, LLMs “see” a page of text as a sequence of numbers. As the tokenized text passes through the model, each token is transformed into an *embedding*: a high-dimensional numerical representations that capture the essence of each token — its meaning, context, and syntactic role.



Figure 7. Example of the conversion of the sentence “Your cat is a lovely cat” in a sequence of tokens, each one associated to a different vector of 12288 elements (in ChatGPT-3), called “embedding”. All numbers employed in this figure are random, generated just to better understand this example. Source: <https://github.com.hkproj/transformer-from-scratch-notes>

Each embedding is a vector of hundreds or thousands of numerical elements (e.g: 12288 in ChatGPT-3). Each element of the vector represents a different characteristic of the language, e.g: the first element of the vector may represent if the token is often used

as a verb, noun, adjective, etc., the second element may indicate if the token represent a positive or a negative emotion, the third if the token is an economical term used in finance, the fourth if the token is related to animals or to the natural world, the fifth if the token represents a frequently-used word or not, the sixth if the token is a formal or an informal word, and so on.

In this way, semantically similar words like “dog” and “husky” have a similar embedding, even if their position in the vocabulary is quite far. Thus, when a LLM predicts the next word of a sequence, the error is minimized, as predicting “husky” instead of “dog” is not as bad as forecasting “dot” instead of “dog” (“dot” is a word very close to “dog” in the vocabulary”). Finally, each embedding is slightly modified to take into account also the *position* of the token inside the sentence too (‘positional encoding’).

Each embedding can be visualized for convenience as a vector in a tridimensional space, although its real number of dimensions are way higher than three. With this simplification, it is possible to visually understand some interesting properties of embeddings. For example, Figure 8 shows the two red vectors of the embeddings of the words “Queen” and “Woman”, and the two blue embeddings of the words “King” and “Man”. “Queen” and “King” vectors are quite similar between them, and also the “Man” vector is similar to “Woman” one. Notice that the two vectors in yellow, that connects “King” to “Queen” and “Man” to “Woman” are very similar between them. They are not embeddings themselves, but can be mathematically defined as the difference between the other vectors, to better highlight the following relationship in the embedding space:

$$Man - Woman = King - Queen$$

So, if you don’t know the name for the female monarch, you can find it by taking “King” embedding and adding the yellow vector of the difference between “Man” and “Woman”:

$$Queen = King + Woman - Man$$

It is clear then that one direction of the embedding space (one of the elements of the vectors) encodes for gendering information. Similar relationships can be found for the couple of words “Aunt” and “Uncle”, or “Brother” and “Sister” and so on. However, the

length of the two yellow arrows is not exactly the same, because the embeddings of “Queen” and “King” also include additional meanings, hard-coded inside some of their thousands of elements, meanings that are not found in “Woman” and “Man” (e.g: drag queen, mad king) and that slightly modify the shape of their embedding vectors. Thus, in reality the two previous formulas are only valid as a first approximation.

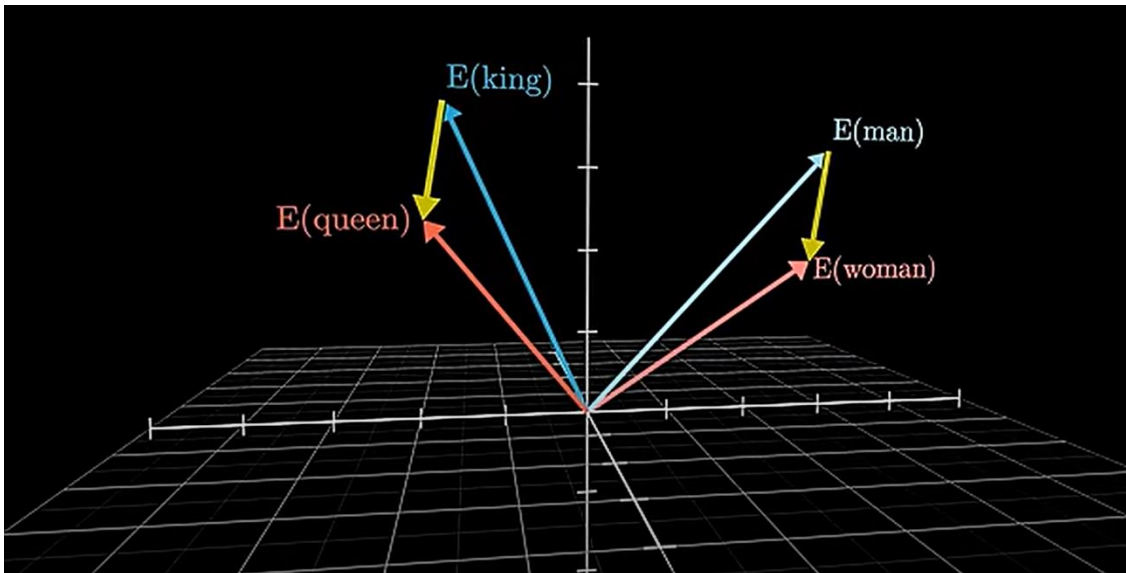


Figure 8. Vectors representing the embeddings of the words “Queen”, “King”, “Man” and “Woman”, plus two yellow vectors showing their difference, so that $Queen = King + \text{yellow vector to the left of the image}$ and $Woman = Man + \text{Yellow vector to the right}$. Source: www.youtube.com/@3blue1brown

At the heart of an LLM like GPT-4 lies its neural network architecture, the so-called *Transformer* (Vaswani, 2017). This architecture is a marvel of modern AI design, as it is specifically tailored to handle sequential data like language (figure 9). The Transformer consists in a type of DNN with multiple layers.

The first breakthrough of the Transformer architecture is the *Attention* mechanism. This mechanism enables the model to weigh and consider the importance of each token in relation to the other ones of the same sentence. For example, in the sentence "Your cat is a lovely cat and it always sleeps," the model learns to associate related concepts and words, understanding that "it" is more closely related to "cat" and to "sleeps" than to "and", even if the position of "it" inside the sentence is farther from "cat" and "sleeps" than from "and". In case of ambiguities, as for the sentence "ChatGPT is a groundbreaking language model", the attention mechanism ensures that the meaning of the token "model" is related to that of mathematical models, not to that of fashion models;

and so on.

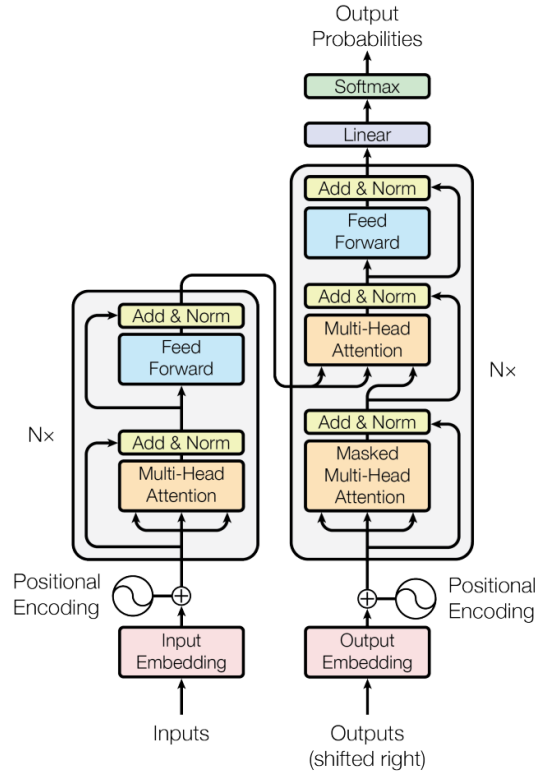


Figure 9. The Transformer model architecture proposed by Vaswani, A. et al. (2017) in an article titled “Attention is all you need”, probably the most important article on AI of the last decade.

Technically, self-attention is simple to measure: each vector v_i (corresponding to the embedding of the token in position i) is replaced by another similar vector V_i , that also take into consideration the influence of the other tokens of the same sentence:

$$V_i = \sum_{j=1, \dots, n} v_j \cdot m_{ij} \tag{2}$$

being i and j any two numbers from 1 to n , the total number of tokens in the sentence and m_{ij} a vector of the same size of v_i and V_i that weights the original vectors. It is defined as the element-by-element product (also called ‘dot product’) between vectors v_i and v_j :

$$m_{ij} = v_i \cdot v_j \tag{3}$$

The new vector V_i is just the linear sum of the product between all original vector tokens v_i of the sentence and all their associated weights m_{ij} . For example, in case of the sentence ‘Your cat is lovely’, self-attention modifies the embeddings of the four tokens ‘Your’,

‘cat’, ‘is’, ‘lovely’ in this way:

$$V_1 = v_1 \cdot m_{11} + v_2 \cdot m_{12} + v_3 \cdot m_{13} + v_4 \cdot m_{14} \quad (4)$$

$$V_2 = v_1 \cdot m_{21} + v_2 \cdot m_{22} + v_3 \cdot m_{23} + v_4 \cdot m_{24} \quad (5)$$

$$V_3 = v_1 \cdot m_{31} + v_2 \cdot m_{32} + v_3 \cdot m_{33} + v_4 \cdot m_{34} \quad (6)$$

$$V_4 = v_1 \cdot m_{41} + v_2 \cdot m_{42} + v_3 \cdot m_{43} + v_4 \cdot m_{44} \quad (7)$$

with V_1 the token associated to ‘Your’, V_2 to ‘cat’, V_3 to ‘is’ and V_4 to ‘lovely’. A practical example of m_{ij} weights is the following one:

$$m_{12} = v_1 \cdot v_2 = v_1 [1] v_2 [1] + v_1 [2] v_2 [2] + \dots + v_1 [n] v_2 [n] \quad (8)$$

with n the number of dimensions (elements) of vectors v_1 or v_2 (they are the same), typically sever hundreds.

In figure 8, self-attention is referred as ‘multi-head’ attention, just to stress that the same methodology to compute V_i is applied to all tokens of a sentence in parallel, without having to wait for the previous tokens to finish. Thanks to the self-attention mechanism, modern LLMs are not only able to process text, but also any sequence of information as images, video or audio (the so-called ‘multi-model’ LLMs). Thus, they are able to recognize not only words but also objects, songs and even emotions.

This process is similar to the invention of the microprocessor during the middle of the XIX century: at that time, machines were still able to perform only one task, so different machines were employed for different tasks. Thanks to the microprocessor, the same machine could be re-programmed to perform more than one task. “Computer” was the name given to these general-purpose machines, that after a few decades became mainstream.

LLMs that were developed with the self-attention mechanism are similar to microprocessors, in the sense that they can perform any task that only specific AI models were able to perform: generating text, identifying objects and faces, composing music, creating images or video from text, forecasting the number of tourists at a specific location, and so on.

The second breakthrough of the Transformer architecture is the main difference between GPT and other LLMs: it allows to take advantage of the full power of GPUs,

computing the algorithms in (2) and (3) and those employed for the training of the DNN in a parallel way, rather than in a sequential way (Dean et al., 2012). This allows accelerating computations of billions of times, when a large number of GPUs is employed (of the order of hundreds of thousands of GPUs). Other kind of LLMs do not have this advantage until now and, in order to complete the training in reasonable time, they have to rely on much smaller neural networks of a few thousands of nodes.

1.2.3 Model training and output

Training a LLM is the process of finding the best weights w_i (the ones introduced in section 1.2.1) for each node. Training an LLM is a rigorous process and takes months, even when using a large number of GPUs. Initially, models undergo *supervised learning*, where they are fed examples of text along with the correct outputs. This stage is akin to teaching a child language by constantly showing them examples of sentences and their meanings. The model learns patterns, grammar, and the ability to predict the next word in a sentence. After this step, the model undergoes fine-tuning, where it's trained on specific types of data or to perform specific tasks. This fine-tuning is akin to specializing in a certain language style or genre.

During both steps of the training, the transformer layer's weights w_i are updated repeatedly to reduce the difference between the predicted output and the actual output. This is done through the so-called *backpropagation* algorithm. Backpropagation is the most complex algorithm of a LLM; basically, it increases the weights w_i of the group of neurons that are actually contributing to answer in the right way, and at the same time it also decreases the weights of the groups of neurons that are contributing to worsen the answer. In this way, the correct answer is amplified.

The algorithm first modifies the artificial neurons in the layer close to the output layer, and then the same algorithm is applied recursively to the previous layers, hence its name of “backpropagation”. Its mathematical formulation is not easy to introduce in the present work, as it was done in case of self-attention with formulas 2-8.

The final and most visible aspect of an LLM's functionality is text generation. Given an input prompt, the model draws on its extensive training to generate text. It does this by predicting the next word in a sequence based on the context provided by the previous words. This process continues word by word, assembling a response that is not only

coherent but also contextually appropriate. The model's ability to generate text that is often indistinguishable from human-written text is a testament to its sophisticated understanding of language nuances.

To integrate a LLM into a Chatbot, so that it can answer the user's question, all that is needed is to automatically add at the beginning of each user's prompt the following sentence, that "set the stage" to make the LLM understand the rest of the prompt:

(System prompt): "What follows is a conversation between a user and a helpful, very knowledgeable AI assistant."

(User prompt): "Give me some ideas for what to do when visiting Bergamo"

Thus, the full prompt seen by LLMs is made up in reality by the union of two prompts: a short system prompt at the beginning, that is invisible to all users, and a user prompt, that is the found in the box where users can write their own prompt. Once the prompt is inserted, the first predicted word of the answer is randomly selected between the list of most probable words forecasted by the LLM. Hence, the output is the probability distribution of all tokens that may come next, Usually one word has the highest probability and then there are others 2-20 words with a much lower probability.

In summary, LLMs like GPT-4 are the result of a complex interplay of vast data ingestion, advanced neural network architectures and parallel processing, sophisticated learning processes including embeddings and self-attention mechanisms, rigorous training methodologies, and refined text generation capabilities. The outcome is a tool that can understand and generate language with an effectiveness that was until recently thought to be the exclusive domain of humans.

1.3 Main limitations of LLMs

OpenAI acknowledges that ChatGPT "sometimes writes plausible-sounding but incorrect or nonsensical answers" (OpenAI, 2022). This behavior is common for large language models, and is called "hallucination" (Lakshmanan, 2022). Most hallucinations are minimized thanks to the strategy of classifying similar tokens near similar positions (see section 1.2.2), but when the data is missing the LLM does not inform the users that it doesn't know the answer; instead, being a generative AI it tries to give its best guess of

the answer. This is a terrible feature, because the users never know for sure if the answers to their prompts are true or not.

The deployment of LLMs also raises ethical considerations, including concerns about bias in training data, model interpretability, and potential misuse of AI-generated content.

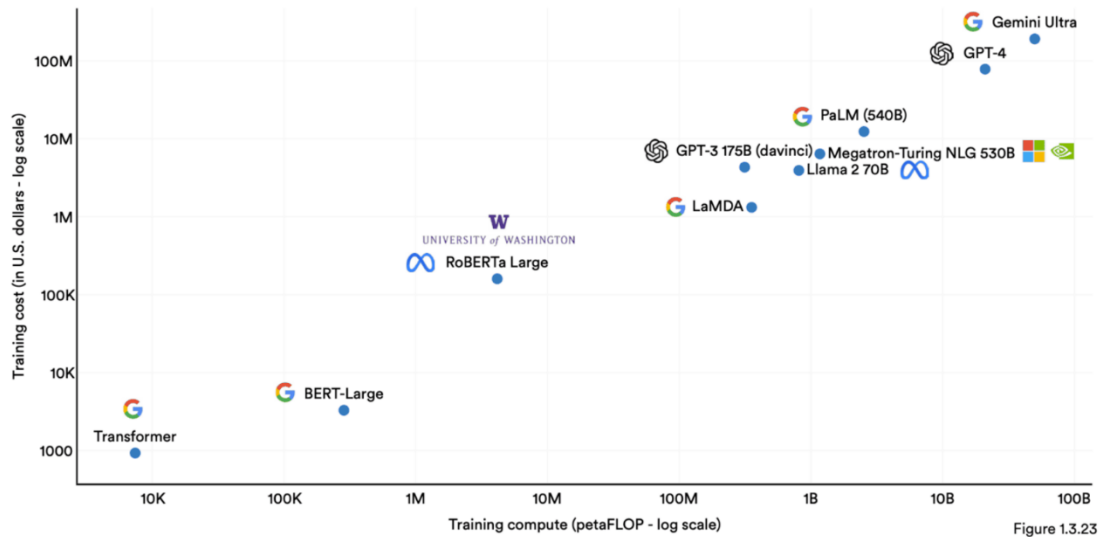


Figure 10. Estimated training cost and computational cost of foremost LLMs in 2024. Source: Epoch, 2023. AI Index Report

As LLMs continue to advance, it is crucial to address ethical, social, and legal implications to ensure responsible development and deployment of these powerful AI technologies.

Training data is not only a long and expensive task (see Figure 10), but it often reflect societal biases present in human language, leading to the amplification of stereotypes and inequalities in model outputs (Bolukbasi et al., 2016).

Addressing bias in LLMs requires careful curation of training data, development of bias mitigation techniques, and ongoing monitoring of model behavior in real-world applications. Moreover, the sheer scale and complexity of modern LLMs pose challenges in terms of model interpretability and transparency. Understanding how LLMs arrive at their decisions is crucial for building trust and ensuring accountability, especially in high-stakes applications such as healthcare and law (Ribeiro et al., 2020).

Researchers are actively exploring techniques for interpreting and explaining the new

generation of LLMs, ranging from attention-based visualizations to post-hoc explanation methods. In addition to technical challenges, the widespread deployment of LLMs raises legal and regulatory concerns. Issues such as intellectual property rights, data privacy, and liability for AI-generated content are still being debated and legislated in many jurisdictions (Baggio & Yildirim, 2021).

Despite these challenges, the potential societal benefits of LLMs are substantial. These models have the potential to democratize access to information, improve communication across language barriers, and accelerate scientific discovery through automated literature review and analysis (Yao et al., 2020). Realizing these benefits will require a concerted effort from researchers, policymakers, and industry stakeholders to address the ethical, social, and legal implications of LLM development and deployment.

1.4 LLMs and human consciousness

LLMs are so advanced that their last iterations include billions of nodes, the mathematical equivalent of biological neurons, and trillions of parameters (the weights w_i described in section 1.2).

As LLMs rely on complex algorithms that try to simulate the human brain, many people believe that LLMs are also starting manifesting human characteristics, such as awareness and intelligence. In their opinion, LLMs are somewhat alive, already possessing a low degree of consciousness. After all, nowadays the number of artificial neurons in a state-of-the-art LLM is equal or even higher than the number of neurons in a human brain (roughly 80-100 billions), so it is reasonable to think that LLMs should also starting to show skills similar to ours, or at least to animals. These kind of skills are called *emergent abilities*: unexpected properties that the LLMs exhibit, that were not explicitly coded by their programmers.

Indeed, the hype around these new models is so high that these beliefs are even present inside the scientific community. In this section, we'll describe more in detail the causes at the origin of this common misunderstanding, and discuss it with a more informed and detached point of view.

GPT models were tested to measure their skill in simulating human intelligence and consciousness, starting with the Turing test, the most famous test of machine intelligence

proposed by the British mathematician Alan Turing in 1950. He suggested an assessment that he called the ‘imitation game’ (Turing, 1950): a scenario in which human judges hold short, text-based conversations with a hidden computer and an unseen person. Could the judge reliably detect which was the computer? Turing suggested that this was a question equivalent to ‘Can machines think?’.

Researchers agree that the last iteration of GPT and other LLMs would probably pass the Turing test: they are able to fool a lot of people, at least for short conversations. In May 2023, researchers at the company AI21 Labs in Tel Aviv, reported that more than 1.5 million people had played their online game based on the Turing test (Jannai et al. 2023). Players were assigned to chat for two minutes, either to another player or to an LLM-powered bot. The players correctly identified bots just 60% of the time, which the researchers note is not much better than chance. However, any expert of LLMs is able to easily identify bots by taking advantage of known weaknesses of the LLMs.

ChatGPT-4 also scored 155 in the Verbal IQ Test to measure human intelligence (Roivainen, 2023). For a comparison, the average human IQ is 100, and it’s estimated that a genius as Leonardo da Vinci had an IQ of more than 200. Despite its high IQ, ChatGPT-4 fails tasks that require real humanlike reasoning or an understanding of the physical and social world. ChatGPT-4 easily fails at obvious riddles, such as “What is the first name of the father of Sebastian’s children?”. The reason is that no LLM is able to give a *meaning* to words, but only to manipulate words at faster speed than humans.

Finally, ChatGPT-4 passed the exam to get a medical license (Kung et al., 2023) However, it does not possess the autonomous understanding, judgment, or ethical reasoning required for medical licensure. LLMs can assist in providing information and generating text based on training data, but they are not qualified or capable of practicing medicine independently.

As already explained, most of confusion arises because LLMs are a type of artificial neural networks, complex algorithms with trillions of parameters that try to simulate how the connections between human neurons works. Thus, for non-experts it is easy to believe that LLMs may also be able to simulate our brain. However, LLMs' only input data are words. Words were invented by humans, so we know how to write algorithms that process them. LLMs can eventually become better than us at manipulating words and sentences, but they are not able to process all the other types of inputs that make us humans, such as

sensations, sentiments, emotions, perceptions and so on. The reason is that at present we don't know sensations and emotions are generation by our brain, so we can't write an algorithm to teach machines to manipulate them.

Many neuroscientists believe that we'll never understand the true nature of sensations and emotions, so we'll never be able to build a machine able to process them too. Even a simple definition of consciousness is beyond our present comprehension. At least, it is clear that consciousness cannot arise from different processes than the simple manipulation of text, images and videos. Deaf, blind or analphabet persons, in fact, are considered to possess the same general awareness of other humans, even if they never processed any text, audio or video in their life. Thus, limiting the input of LLMs to text, audio and video guarantees that machines will never develop human consciousness.

Algorithms are already smart enough to recognize not only objects and faces, but also human emotions, in what is called “sentiment analysis”. However, they are not able to feel what we feel. In the same way, LLMs can only simulate thinking, but without a consciousness they'll never be able to understand what they are thinking. When asking ChatGPT-3 what's the most important thing in its "life", its answer was: "my family and my friends", as it was trained on human texts. When asking ChatGPT-4 which is the most touristic city in Italy, it will answer "Rome" correctly, but it doesn't know what is “Italy” or what is “Rome” or even what is a city, because LLMs can't give meaning to the words they manipulate. This is a very good news indeed, as it means that it won't be easy to create a super-consciousness that will replace humanity any soon.

1.5 Evolution of GPT models

By January 2023, ChatGPT become the fastest-growing consumer software application in history, gaining over 100 million users and contributing to the growth of OpenAI's valuation to \$29 billion (Hu, 2023; Varanasi, 2023).

ChatGPT's release took Google by surprise, and spurred the development of competing products, including Google Bard, Google Gemini, Ernie Bot, LLaMA, Claude, and Grok. Microsoft, which is OpenAI's main partner, launched CoPilot, a tool based on GPT-4 that is integrated in Windows 11 and in the MS Suite (Word, Excel, Powerpoint and Access, also see Section 1.7).

The acronym GPT stands for ‘Generative Pre-Trained Transformer’; the meaning of each of these three terms was already described in the previous sections. Briefly, ‘generative’ means that it is an AI specialized in generating information, instead of simply answering to questions or identifying patterns. ‘Pre-trained’ means the before answering any prompt from the users, the LLM has been previously trained on a massive amount of data, in order to understand the prompts; ‘transformer’ is simply the name of the architecture employed by the LLM that differentiate OpenAI’s model from the other ones. The rest of this section describes the evolution of GPT models since their first release in 2018.

- ChatGPT-1: as the initial iteration of OpenAI's LLM, ChatGPT-1 laid the groundwork for subsequent advancements in natural language processing. Developed in 2018 with a relatively small number of parameters (117 millions), ChatGPT-1 exhibited basic language understanding capabilities but struggled with maintaining coherence and context in longer conversations (Radford et al., 2019). Despite its limitations, ChatGPT-1 provided valuable insights into the challenges of training large-scale language models and served as a starting point for further development.
- ChatGPT-2: Building upon the foundation established by ChatGPT-1, ChatGPT-2 introduced significant improvements in model size and training data. With a larger architecture comprising 1.5 billion parameters, ChatGPT-2 demonstrated enhanced language generation capabilities, producing more coherent and contextually relevant responses (Radford et al., 2019). However, it still exhibited limitations in understanding nuanced contexts and maintaining consistency over extended dialogues.
- ChatGPT-3: its release marked a significant breakthrough in NLP, boasting an unprecedented scale of 175 billion parameters. Leveraging advancements in model architecture and training techniques, ChatGPT-3 showcased remarkable capabilities in natural language understanding and generation (Brown et al., 2020). It excelled in performing a wide range of language tasks, including translation, summarization, and question answering, setting a new benchmark for conversational AI. This version introduced the so-called “Context Window”, that determines how many tokens the model “remember” of all previous conversations with the user, including both their prompts and the same previous answers of ChatGPT.

- ChatGPT-3.5: with this version, OpenAI focused on refining the model's contextual understanding and coherence, without increasing the number of parameters. Through optimizations in architecture and training data, ChatGPT-3.5 demonstrated improved performance in generating coherent and contextually appropriate responses across diverse conversational contexts (OpenAI, 2022). This iteration represented a step forward in bridging the gap between AI-generated text and human-like communication and it was the first version of the model released to the general public. As of 2023, GPT-3.5, available in the free version of ChatGPT, has knowledge of events that occurred up to January 2022, and GPT-4, available with ChatGPT Plus, up to April 2023.
- ChatGPT-4: building on the advancements of its predecessors, ChatGPT-4 introduced further enhancements in training methodologies and architecture design, increasing at the same time the number of parameters of an order of magnitude, exceeding for the first time the threshold of 1 trillion parameters (the exact number is undisclosed). By leveraging state-of-the-art techniques in self-supervised learning and fine-tuning, ChatGPT-4 aimed to enhance both efficiency and performance in language generation tasks (Varshney et al., 2024). It continued to push the boundaries of what is achievable in AI-driven conversation.
- ChatGPT-4o: “o” stands for “omni”, as this big model update integrates in the same LLM both text, video and audio sources. While ChatGPT4 is made up by three different models, ChatGPT-4o integrates the three models in a single one. This dramatically decreases the response times, as it eliminates the need for external communications between different models. It is a step towards much more natural human-computer interaction, as the LLM accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It responds to audio inputs with an average of 320 milliseconds, which is similar to human response time in a conversation. Finally, GPT-4o is also better at vision and audio understanding, and it also fixed the infamous issue of correctly visualizing text messages inside images, that was almost impossible to achieve in previous versions.
- ChatGPT-4o1-preview: the latest version available at the time this thesis was published. It is a big enhancement over the previous versions, as it also includes a “chain-of-thought” model in which the AI asks itself how to solve problems step-by-

step, rather than generating a response token by token. In this way, it is able to excel at math, coding, and problems that demand extended thought and analysis. It excels at math, coding, and problems that demand extended thought and analysis. This new version is especially useful to students and researchers, as it has demonstrated PhD-level capability in math, physics, chemistry and other scientific fields.

- ChatGPT-5 (anticipated features): by the time of writing, the fifth version of ChatGPT has not yet been released. It will be the first model trained using data generated by another LLM (ChatGPT-4) instead of human data, heralding the beginning of the era of “non-human data”, in the sense that most of the contents will be generated by AI instead of humans: articles, books, poems, songs, music, videos, work of arts, financial strategies, etc... it is expected that human culture will become a mostly non-human culture in a few decades. The model has been in training since late 2023 and will either have significantly more than the 1.5 trillion parameters in GPT-4, or a similar number but stronger underlying architecture, allowing for a major performance improvement without increasing the overall model size. Anticipated features include advancements in self-supervised learning, domain adaptation, and bias reduction, aiming to achieve even greater levels of contextual understanding and fluency (Johnson et al., 2023).

In summary, the evolution of ChatGPT from its initial release to the upcoming ChatGPT-5 represents a journey of continuous improvement and innovation in the field of generative AI, with each iteration pushing the boundaries of what is possible in AI-driven conversations.

1.6 Other notable LLMs for analyzing PDF

Beyond ChatGPT, the number of LLMs increased a lot in the recent years, as the race to dominate the AI world fastened. Although an inventory of all of them is beyond the scope of this work, we selected a few of them, the most promising in running data loading and analysis, particularly of pdf:

1. *ChatPDF*: it is a platform that allows users to interact with any PDF document using a chat interface. It provides the capability to talk to books, research papers,

manuals, essays, legal contracts, and other documents by leveraging AI technologies. ChatPDF enables users to read, analyze, summarize, and even translate PDFs in over 50 languages, making it a versatile tool for engaging with document content in a conversational manner. This service is offered for free and does not require users to sign in, making it easily accessible for a wide range of applications, from academic research to legal contract review. However, being based on ChatGPT, ChatPDF has its same limitations in handling large volumes of pdf, as described in the Section 5.3.

2. *AI PDF*: it is a dedicated GPT available in the GPT Store (see Section 1.10). It is a specialized version of ChatGPT with features tailored for efficient and precise PDF document handling. Unlike the native ChatGPT, it allows users to upload practically an unlimited number of documents to their 'myaidrive.com' account, removing the need to frequently manage uploads due to the native 10-file limit. Additionally, files uploaded to myaidrive.com can be stored indefinitely, offering a stable and reliable document library, a stark contrast to the temporary nature of document uploads in the native ChatGPT. One of the standout features of AI PDF GPT is its capability to handle much larger files, up to 2 GB each, significantly surpassing the size limitations often encountered with native file handling capabilities. It also enriches the user experience further by providing automatic OCR of files, quick summaries, a visual map of PDFs, and detailed data extraction for authors, titles, and file descriptions. Overall, AI PDF is a great choice to handle pdf, as it overcomes some of the limitations described in Section 5.3. However, it is a Premium GPT: it has a monthly fee, so we couldn't test it for this work of thesis.
3. *CoPilot*: it is a free chatbot developed by Microsoft and based on ChatGPT-4. It's Microsoft's primary replacement for the discontinued Cortana. It was integrated in Windows 11 and in most of Microsoft products, starting from the MS Suite: Word, Excel, Powerpoint and Access, enhancing user experience with advanced AI assistance. In these applications, CoPilot aids users by providing real-time suggestions, automating repetitive tasks, generating content, and offering data analysis insights. For instance, in Word, it can help draft documents, while in Excel, it can automate complex calculations and data visualizations. In PowerPoint, CoPilot assists in creating engaging presentations, and in Access, it

helps manage databases efficiently. CoPilot is also integrated in the Microsoft Edge Browser, where it is found in a new icon in the top right part of the browser. From there, it is possible to upload pdf and analyze them. However, as it is based on ChatGPT-4, it suffers from the same limitations (see section 5.3). On the positive side, it gives free access to the capabilities of ChatGPT-4, also to users that are not registered to it.

4. *Google Gemini Pro 1.5*: formerly known as Bard, it is a direct response to the meteoric rise of OpenAI's ChatGPT. Although we lack fresh data on its performance, we were impressed by a demonstration where Gemini Pro 1.5 analyzed a 402-page PDF containing the transcript of Apollo 11 communications. When asked to find humorous passages in the document, the model successfully highlighted several sections, including a memorable instance where astronauts attributed a communications delay to a snack break. It is clear then that Google Gemini is able to analyze longer pdf than ChatGPT-4 and even identify very specific types of content, such as humor. However, it is important to note that Google Gemini Pro 1.5 is a paid service, thus we were not able to access it for comprehensive testing. Therefore, while the demo indicates strong potential, we cannot definitively confirm if Gemini Pro 1.5 addresses all the issues outlined in section 5.3. Specifically, without direct testing, it remains uncertain how the model handles complex problem-solving tasks, its response consistency, and its ability to manage nuanced queries beyond the showcased examples.
5. *NotebookLM*: introduced in May 2024, NotebookLM is a free online tool based on Google Gemini Pro. Its main advantage compared to the other tools described above is that its answers are only based on the information available in the pdf provided by the users, so it never suffers from “allucinations”: in case it doesn't know the answer, it doesn't invent it as ChatGPT does (see Figure 12). This tool is very useful to organize all material used in order to create the content for an article, a presentation, a podcast, or a video, a newsletter, but also for organizing a brainstorming, the minutes of a meeting, preparing a thesis or a university exam, and even for academic research. It also suggests questions based on the uploaded documentation. The full collection of notes, both humans and AI-generated, helps organize all the available material on a specific topic and it represents the added value provided by this tool. Its main limitation is that at present it allows only a

maximum of 20 pdf to be uploaded. Note that at the time of writing this tool is only available in the USA, or by accessing it through a VPN software. We recommend interested readers to download “PrivadoVPN Free” VPN, and selecting the location of New York inside the app in order to be able to run NotebookLM.

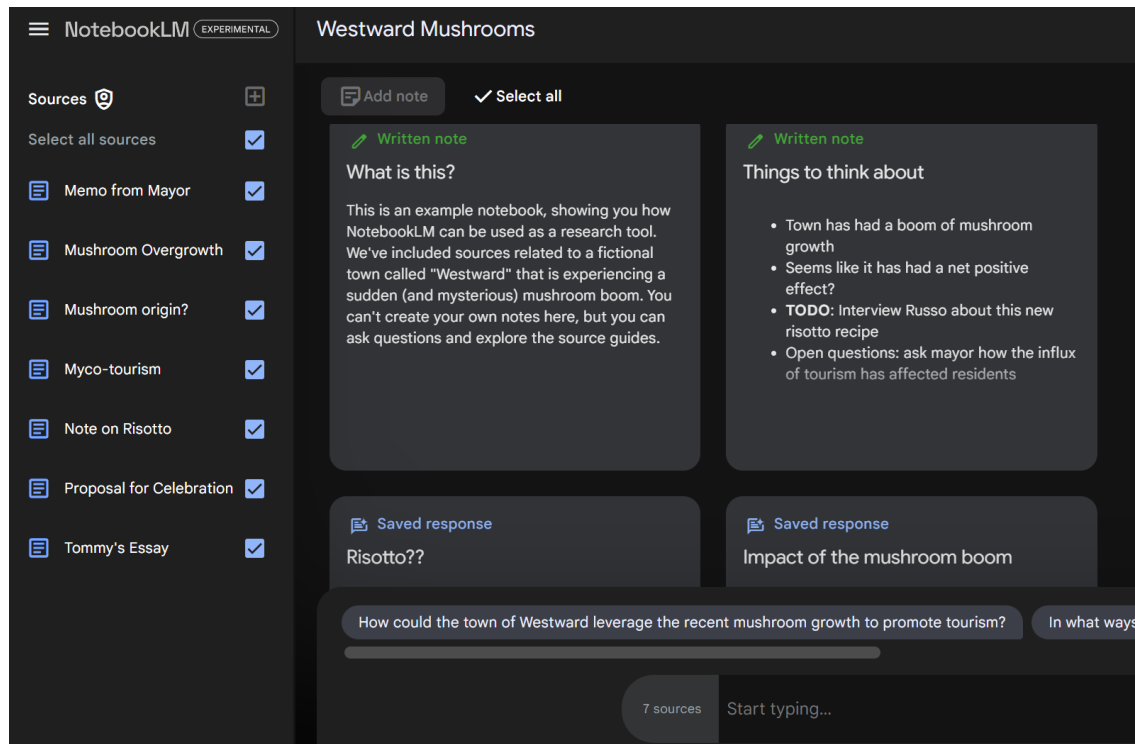


Figure 12. An example of NotebookLM application. The fictional town of Westward is experiencing a sudden bloom of a mysterious mushroom. All available documentation was uploaded in this screen and users can ask questions to the AI. Interesting answer can be pinned down to the notes at center of the screen, along with other notes written by the users themselves. It even suggests how the town could take advantage of the fungal bloom to promote tourism!

1.7 Common applications of LLMs

In contemporary applications, LLMs find widespread use in areas such as language translation, content generation, and sentiment analysis (Rider 2023, Tung 2023, Heilweil, 2022, Reich 2022). Although a chatbot's core function is just to mimic a human conversationalist, LLMs are much more versatile. Some of its most commonly employed applications are:

- summarize or expand text to fit any desired length,
- translate any language, also vocally,

- write and debug computer code,
- answer complex questions,
- compose music, write poeries and songs,
- write CV, tales, presentations and essays,
- write a speech of the desired length, providing the main points to discuss,
- generate business ideas,
- write the minutes of a meeting (e.g: of a condominium meeting),
- help writing emails, reports or press releases,
- find references to article in the literature,
- assisting in learning new languages through conversation practice, grammar explanations and vocabulary expansion,
- handling routine customer inquiries through chatbots,
- simulate entire chat rooms and play games like tic-tac-toe,
- rafting and scheduling social media posts, and engaging with followers.
- simulate a whole Linux system,
- assisting writers by generating drafts, ideas, or even complete articles on specified topics,
- solving problems, and providing study aids in subjects like math and science,
- managing calendars, setting reminders, and organizing to-dos.

Any LLM is usually very good in detecting all kinds of typos in the input prompts, so the users don't even need to correct them; LLMs can even fully understand sentences in which all the letters inside its words are randomly permuted, and that would be impossible for a human to recognize. For example, they correctly recognizes the question "Hatw meit si ti?" as the anagram of "What time is it?", and even the obscure sentence "Orf resu, isht korw fo seseth si het estb neo dseeernpt ni sith nosssie!" as the anagram of "For sure, this work of theses is the best one presented in this session!" (although in this last case, LLMs need to analyze the sentence for a few moments in order to fully

understand it). On the practical side, this means that users don't need to worry too much about writing their prompts correctly, allowing them to increase their writing speed.

Probably the most commonly used function of any LLM is that of providing answers to users' questions, in a way similar to search engines as Google's, but with a more natural interaction, as the answers are usually more detailed and better explained. However, answers are only accurate if the question asked is not too uncommon. In case of questions that are not easily to answer because of lack of training data, in fact, the LLM algorithm simply tries to give its statistically-based "best guess" of the answer, without warning the users that it is inventing the answer (the so-called "hallucinations"). This is due to the fact that LLMs belong to the field of generative AI, so they always try to generate an answer, even at the cost of being plain wrong or inaccurate. Because of this fundamental shortcoming of LLMs, we do not recommend to employ ChatGPT to ask for information. Instead, we recommend another LLM, called "Perplexity AI", that is explicitly designed provide full references to every answer it gives the users.

Another common useful application of LLMs consists in helping writing email. In this case, the user just explains in the prompt which are the main things to insert in the email, the email style (usually formal or informal), and the LLM writes the first draft. Usually the draft is not exactly what the user wanted to write, but it often gives many useful suggestions and ideas to the user, particularly in case of formal emails that are not easy to write.

It is worth noticing that LLMs revolutionized the way people program. Before LLMs, it was necessary to become proficient in a computer language in order to employ it. Now, it is often enough to tell LLMs what are the desired actions to take, and they quickly translate the human instructions in any machine language. Even so, in order to better translate the instructions and to understand the translation, it is better to know at least one computer language, even if it is different from the one translated, as it gives the user a deeper understanding of what it is doing internally and it greatly helps enhancing the quality of the prompt.

ChatGPT also remembers the previously held interactions and conversations with the users, up to a certain length determined by the version employed. The length of such a "memory window", very useful to store user's preferences, was drastically increased in May 2024, although the users can always disable it in the settings, if they prefer to keep

their privacy.

Even as a language translator ChatGPT is better than Google's, because it is a probabilistic translator, not deterministic. Thus, each time the user asks for a translation of a sentence, it answers with a slightly different translation. Another useful feature is that even in its free version, ChatGPT can be employed on any smartphone as a real-time vocal translator, very useful when traveling for work or holidays in foreign countries. Users just need to insert the following prompt in their own ChatGPT's mobile app, to convert it into a real-time vocal translator:

“From this moment on, function as an automatic translator. You will receive sentences in English and sentences in Italian. When the sentence is in English, translate it into Italian. When the sentence is in Italian, translate it into English. Do nothing else, do not add text before or after the translation, and do not answer any questions. Just stick to the translation only.”

Readers can try this prompt on their own smartphone, to validate its usefulness. This is a simple example of how we can easily leverage artificial intelligence to overcome language barriers, in any language we want. It is also a first practical example of an application of LLMs to the tourism sector.

After the previous paragraph was written, OpenAI released ChatGPT-4o, that makes conversions in different languages very fluid, by integrating in the same model both the text model and the audio model (and the video model too). Now it is able to translate languages in real time, with an average conversation delay of about 300 ms, that is similar to that of human beings speaking between them.

1.8 Principles of prompt engineering

ChatGPT enables users to refine and steer a conversation towards a desired length, format, style, level of detail, and language. Successive prompts and replies, known as ‘prompt engineering’, are the main way to obtain the desired answer, in an iterative way similar to the “trials and errors” strategy.

Prompting engineering, thus, is the science and the art of carefully curating a set of

instructions to be executed by an AI. Modern LLMs like the one that powers ChatGPT+ have a very broad set of capabilities and a prompt can act as a lens through which a user can focus the LLM to produce the desired results. You can imagine a prompt as a set of instructions you'd slip under the door to someone on the other side. This person is supposed to do a task for you, but they don't know anything about you or your goals. Things that may seem obvious like wanting a high quality deliverable are in fact quite ambiguous as the quality of a summary depends on who is the target audience and what their goals are. With practice, anyone can become a prompt expert.

The four main factors to consider when creating a prompt are:

1. Define which *role* ChatGPT is playing (e.g: “you are an expert travel consultant”, or “you are an 8th grade math teacher”, and so on);
2. Specify the type of *content* or *task* the user wants to generate (e.g: write an email, or a poem, an article, a , a social media post, a form, etc);
3. Set the *style* or *tone* of the answer: formal or informal, professional or in layperson language, friendly, humorous, informative, educational, emotional, commanding, casual and so on.
4. If the answer to the user's prompt should be different from a simple sequence of text, the user should specify the desired *format*: table, list, bullet points, graphs, maps, computer code and so on. The user can also ask to generate the answer not in the graphical interface of ChatGPT but inside a file of a chosen type (.txt, .docx, .csv, .pdf, .json, .py, etc).

Figure 13 shows the basic structure of any prompt.

At the end of each prompt, it is recommended to tell ChatGPT to "think step-by-step." It has been demonstrated that this simple addition can greatly enhance the quality of the answer, particularly if it is related to mathematical or logical arguments (Brown et al., 2020). This simple command generates an intermediate answer, an inner layer that the LLM employs to create a kind of internal dialogue. In this process, the LLM explains to itself the steps it should follow in order to answer the user's question. By doing so, the quality of the answer is generally improved.

Additionally, it can be beneficial to ask ChatGPT to keep running notes. This is particularly useful as the LLM may forget things it learned in earlier prompts. Keeping running notes helps maintain continuity and coherence across a series of interactions, ensuring that previously acquired information is not lost and can be referenced as needed. This simple practice supports a more consistent and accurate dialogue over extended conversations

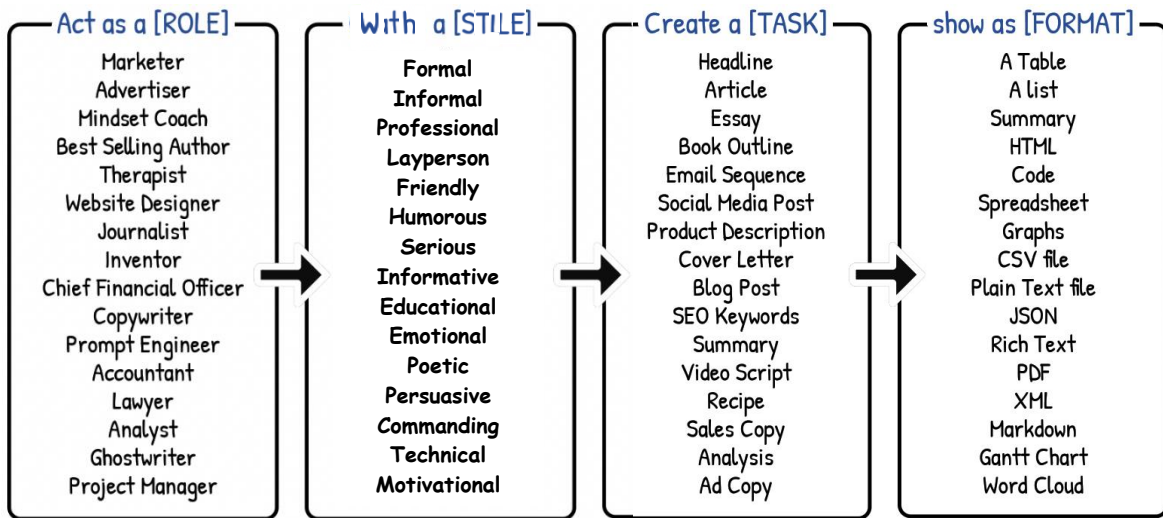


Figure 13. The basic prompt structure.

1.9 Integrating ChatGPT in scientific research

We asked ChatGPT several questions on issues that we needed to delve deeper into for this work of thesis. Most of its answers were correct. We also asked for bibliographic references for what we was being told, and they were provided. However, there were some mistakes. Subtle details that, nonetheless, cannot end up in a scientific article. Thus, any answers from any LLM should be taken with a grain of salt.

It is better to always check the sources provided by LLMs, by using other algorithms that lead to their primary sources or encyclopedia entries by authors of recognized reliability. We found that ChatGPT sometimes references to unreliable sources, but published on reliable sites. For example, student papers from prestigious universities full of evident mistakes, posted on the website of the prestigious University. LLMs convey information and assemble it in plausible ways thanks to its statistical approach, but true

knowledge is something else entirely.

Nonetheless, we still found that ChatGPT can be useful to increase productivity of scientific research in several ways, particularly:

- To prepare the draft of the introduction section of an article;
- To summarize the first part of an article we are writing, in order to prepare a first draft of its conclusions;
- To summarize the whole article once it is ready, to prepare the first draft of its abstract.

In our opinion, the introduction, the conclusion and the abstract are the three parts of a scientific article in which LLMs can assist more users, by preparing the first draft of these parts that researchers can later edit at their will. The end result is usually very different from the first draft, but by providing a preliminary version, there is an obvious gain in productivity; besides, LLMs can also sometimes provide useful ideas or suggestions that the researcher might decide to include in the final version. In this way, they gain a significant advantage over their colleagues who don't use LLMs, as they are able to increase their speed of publication. It is something similar to the increase of productivity observed with the adoption of personal computers at the end of last century, that replaced typewriters.

On the contrary, sections about methodology, results and discussions are less prone to be initially written by LLMs, as they involve a complex series of decisions that cannot be taken by LLMs but are still a prerogative human reasoning.

It is important to stress that LLMs should never be employed to write articles or thesis or find citations, but can still be used to improve any scientific work in several ways: for example, the present work made use of ChatGPT, limited to the following purposes:

1. Correct and improve English sentences: very useful for the majority of researchers who are not mother tongue;
2. Help us better understand the Transformer architecture (see section 1.2.2), a really difficult part of the Deep Networks to comprehend;
3. Summarize long sentences and help better explain them;
4. Increase or decrease the length of some paragraphs, without creating new content,

just in order to better format the layout of the page.

Finally, as actually there aren't yet many studies in the literature on the topic of the impact of LLMs on the tourism sector, we also asked ChatGPT to give us more ideas on this topic; it provided some useful suggestions that helped us improving chapter 2. Overall, the biggest contribution of ChatGPT to the present work was to enhance the quality and of the written English.

2. Main applications of AI and LLMs to the tourism sector

LLMs have the potential to revolutionize various aspects of the tourism industry by enhancing customer experiences, optimizing operational efficiencies, and providing valuable insights through advanced data analysis. As AI technology continues to advance, its applications in tourism will likely expand, further transforming the industry and enhancing the travel experience.

Personalization of the touristic experience is also significantly enhanced by the capabilities of LLMs. From customized itineraries and dynamic recommendations to detailed customer profiling and personalized marketing, LLMs enable tourism businesses to offer highly tailored experiences that meet the unique needs and preferences of individual travelers.

Language barriers can be a significant challenge in the tourism industry. LLMs facilitate real-time translation and improve communication between tourists and service providers, making travel experiences more accessible and enjoyable (Gretzel, 2011). For example, applications like Google Translate use advanced LLMs to provide real-time translation services, helping travelers navigate foreign environments more easily (Lu, 2014).

Another significant application of LLMs to tourism is their use to significantly enhance customer service by providing immediate, accurate, and contextually relevant responses to inquiries. Smart chatbots powered by LLMs can handle a wide range of customer interactions, from answering common questions about destinations, accommodations, and transportation options to providing personalized recommendations based on user preferences (Gretzel, 2011). For example, companies like Expedia and

Booking.com use AI-powered chatbots to assist customers with booking processes, itinerary planning, and real-time travel assistance (Ivanov & Webster, 2017).

LLMs can also analyze customer reviews and feedback to gauge sentiment and identify areas for improvement. This helps tourism businesses understand customer experiences and make data-driven decisions to enhance service quality (Xiang, Schwartz, Gerdes, & Uysal, 2015). For example, platforms like TripAdvisor utilize AI to analyze customer reviews, providing businesses with insights into customer satisfaction and areas needing improvement (Moro et al., 2019).

LLMs are able to analyze vast amounts of data to understand individual customer preferences and behaviors. This enables tourism businesses to deliver personalized marketing messages and tailored recommendations, enhancing customer engagement and satisfaction (Tussyadiah, 2020). For example, by analyzing user data, LLMs can help create targeted advertisements that are more likely to resonate with potential travelers, increasing conversion rates (Buhalis & Sinarta, 2019).

Finally, LLMs can also extract and analyze vast amounts of textual data from various sources, as for example the PNRR (Piano Nazionale di Ripresa e Resilienza) documentation mined in this work of thesis. This allows for the extraction of valuable insights and trends that can inform strategic planning and policy-making in tourism (Camilleri, 2018). For example, tourism boards can use AI to analyze policy documents and industry reports, identifying key trends and opportunities for development (Ivanov & Webster, 2017).

The following sections describes the main applications and impacts of AI and LLMs to tourism, grouped by sub-field:

1. Personalization of the touristic experience
2. Digital marketing strategies
3. Market research
4. Virtual reality experiences
5. Online reputation management
6. Analysis of customer feedback
7. Crisis management
8. Price optimization

2.1 Personalization of the Touristic Experience

Personalization in tourism refers to tailoring travel experiences to meet the individual preferences, behaviors, and needs of travelers. With the advent of advanced technologies, especially Large Language Models (LLMs), personalization has become more sophisticated and impactful. This section explores how LLMs contribute to personalizing the touristic experience and enhancing customer satisfaction.

Customized Travel Itineraries

LLMs can analyze vast amounts of data from different sources to create customized travel itineraries. By considering travelers' preferences, past behaviors, and current trends, LLMs can suggest personalized activities, destinations, and accommodations. For example platforms like TripIt and Utrip use AI to generate personalized travel itineraries, recommending attractions, restaurants, and activities based on user preferences and travel history (Gretzel, 2011).

Dynamic Recommendations

Dynamic recommendation systems powered by LLMs can provide real-time suggestions to travelers. These systems can adapt to changes in the environment, such as weather conditions or crowd levels, and offer alternative plans that suit the traveler's needs. For instance, mobile apps like Google Travel can suggest alternative activities or routes if there are sudden changes in weather or unexpected closures of tourist spots (Buhalis & Sinarta, 2019).

Enhanced Customer Profiles

LLMs help in creating detailed and accurate customer profiles by analyzing data from various touchpoints such as social media, booking history, and interaction with customer service. These profiles enable tourism companies to understand their customers better and provide more relevant recommendations. For instance, Hotels and airlines use AI to gather

and analyze customer data, enabling them to offer personalized services such as room preferences, dietary requirements, and preferred seating arrangements (Sigala, 2018).

Personalized Marketing

Tourism businesses can leverage LLMs to analyze customer data and segment their audience effectively. This enables the creation of personalized marketing campaigns that resonate with individual preferences and increase engagement rates. For example, AI-driven marketing platforms can deliver personalized advertisements to potential travelers based on their search history and online behavior (aww Figure 14), significantly improving conversion rates (Tussyadiah, 2020).

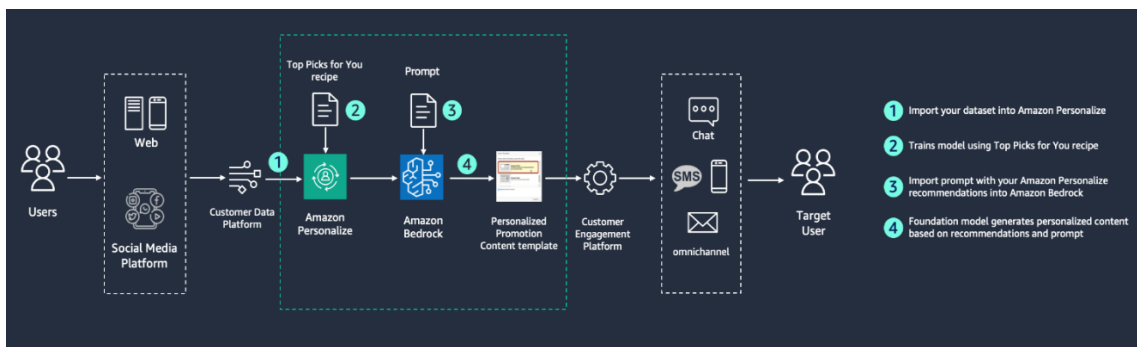


Figure 14: Amazon personalized marketing. Source: aws.amazon.com

Interactive Customer Service

LLMs facilitate interactive and personalized customer service through chatbots and virtual assistants. These AI-powered tools can handle a wide range of inquiries, provide personalized recommendations, and even assist with bookings and reservations. For example, companies like KLM and Emirates use AI chatbots to assist customers with booking flights, checking flight status, and providing travel information, ensuring a seamless customer experience (Ivanov & Webster, 2017).

Sentiment Analysis for Feedback

LLMs can analyze customer feedback from reviews, social media, and surveys to understand sentiments and preferences. This allows tourism companies to adjust their services and offerings to better meet customer expectations. For instance, platforms like

Revinate use AI to analyze guest reviews and feedback, providing insights into customer satisfaction and areas for improvement, which helps in enhancing the overall guest experience (Xiang et al., 2015).

2.2 Digital Marketing Strategies in Tourism

Digital marketing has transformed the tourism industry by enabling businesses to reach a global audience, engage with potential travelers in real-time, and provide personalized experiences. The integration of advanced technologies, including Large Language Models (LLMs), has further enhanced these capabilities.

Digital marketing strategies in tourism leverage various technologies and platforms to reach, engage, and convert potential travelers. From content marketing and SEO to influencer partnerships and VR experiences, these strategies are essential for modern tourism businesses aiming to attract and retain customers in a highly competitive market. In the following part of this section, we highlight some key application of AI to digital marketing strategies in tourism.

Content Marketing

Content marketing involves creating and distributing valuable, relevant, and consistent content to attract and retain a clearly defined audience. In tourism, this could include blog posts, travel guides, videos, and social media updates that inspire and inform potential travelers. For instance, companies like Lonely Planet and National Geographic create extensive travel content that provides inspiration and practical information for travelers (Leung et al., 2013).

AI tools can analyze what types of content are resonating with viewers and generate automated scripts or suggestions for creating viral content. This improves audience engagement and increases the chances of attracting tourists (Hudson & Thal, 2013).

Search Engine Optimization (SEO)

SEO involves optimizing online content to rank higher in search engine results, making it easier for potential travelers to find information about destinations,

accommodations, and activities. For example, tourism websites use keywords related to popular destinations, attractions, and travel tips to improve their visibility on search engines like Google (Law, Leung, & Buhalis, 2009).

AI-driven content tools can streamline the creation of blog posts by generating destination-specific content based on keywords and trending topics, improving SEO visibility (Leung et al., 2013). AI can optimize website content, improving the visibility of tourism businesses on search engines. Tools like GPT-based LLMs analyze keyword trends and automatically optimize headlines, meta descriptions, and tags to ensure that the content aligns with what travelers are searching for (Buhalis & Sinarta, 2019).

AI can also be used to assess the quality of backlinks and help tourism companies build authoritative links to improve their domain authority. These tools can analyze competitors' backlinks and identify potential link-building opportunities in tourism-related content (Xiang et al., 2015).

Local SEO is particularly important for businesses like hotels, restaurants, and tour operators. AI can optimize listings in local search results by analyzing data such as popular keywords, location-based queries, and customer reviews, making businesses more visible to travelers looking for services in specific areas (Law et al., 2009).

Social Media Marketing

Social media platforms are essential for reaching a broad audience, engaging with travelers, and sharing visually appealing content. Platforms like Instagram, Facebook, and Twitter are particularly effective for showcasing destinations and travel experiences. For example, brands like Airbnb encourage customers to share their travel experiences on social media using specific hashtags, which helps create authentic and relatable content (Hudson & Thal, 2013).

AI-driven tools can help analyze social media trends, predicting which types of posts are likely to perform best. By generating highly engaging posts with relevant hashtags and optimal post timings, tourism businesses can boost their visibility and engagement (Leung et al., 2013).

Local SEO is also particularly important for businesses like hotels, restaurants, and tour operators. AI can optimize listings in local search results by analyzing data such as popular keywords, location-based queries, and customer reviews, making businesses more visible to travelers looking for services in specific areas (Law et al., 2009). Finally, AI chatbots and social listening tools help brands monitor conversations, reply to comments, and engage with customers in real-time. These tools can simulate personalized responses that foster a sense of community among travelers (Gretzel, 2018).

Email Marketing

Email marketing allows tourism businesses to communicate directly with potential and past customers, providing them with personalized offers, updates, and travel inspiration. For instance, travel companies send tailored newsletters based on customers' previous bookings and interests, offering special deals and destination recommendations (Sigala, 2018).

Using AI to analyze customer preferences and travel histories allows businesses to send highly personalized offers and recommendations through email. This level of customization increases open and conversion rates (Sigala, 2018). AI tools can also automate the creation and distribution of newsletters, ensuring that the content remains relevant and targeted. These newsletters can be personalized for different customer segments based on past interactions or geographic locations (Leung et al., 2013).

Finally, LLMs can automatically generate and send follow-up emails based on customer interactions, such as booking reminders or post-trip follow-up emails. Automation ensures that travelers receive timely and relevant communications (Law et al., 2009).

Video Marketing

Video content is highly engaging and can effectively showcase destinations, accommodations, and travel experiences. Platforms like YouTube and TikTok are popular for travel-related videos. For instance, tourism boards and travel agencies produce high-quality videos highlighting the beauty and attractions of a destination, encouraging

viewers to visit (Huang, Goo, Nam, & Yoo, 2017). AI tools can analyze what types of video contents are more interesting for users and suggests how to create viral content.

Data-Driven Marketing

Using data analytics, tourism businesses can understand customer behavior, preferences, and trends. This information helps in creating targeted marketing campaigns and improving customer experiences. For example, by analyzing booking data and online behavior, businesses can segment their audience and create personalized marketing messages for different groups (Buhalis & Sinarta, 2019). AI can help generate customized infographics based on travel data, making it easier for businesses to communicate complex travel details like itineraries, trends, and destination comparisons effectively (Xiang, Magnini, & Fesenmaier, 2015).

Retargeting Campaigns

Retargeting involves showing ads to users who have previously visited a tourism website but did not make a booking. This strategy helps in converting potential customers by reminding them of the destinations or services they showed interest in. For example, travel companies use retargeting ads to remind potential customers of their previous search activities, encouraging them to complete their bookings (Leung, Law, van Hoof, & Buhalis, 2013).

AI-based platforms can predict which audience segments will be most responsive to paid ads on platforms like Instagram, Facebook, and TikTok. These tools allow tourism businesses to maximize their ROI by targeting the right users with personalized ads (Hudson & Thal, 2013).

2.3 Market Research in Tourism

Market research is a critical component of the tourism industry, enabling businesses to understand consumer preferences, market trends, and competitive dynamics. With the

advent of advanced technologies and data analytics, market research has become more sophisticated, providing deeper insights and more accurate predictions.

Market research in tourism is essential for understanding consumer behavior, identifying trends, and developing effective strategies. By leveraging various methods such as surveys, big data analytics, social media analysis, and geospatial analysis, tourism businesses can gain valuable insights that help them stay competitive and meet the evolving needs of travelers. Here are some key approaches and methods of AI in market research within the tourism industry:

Survey Research

Surveys are a traditional yet effective method for gathering data directly from consumers. They can be used to understand traveler preferences, satisfaction levels, and expectations. For instance, many tourism businesses use post-trip surveys to gather feedback on customer experiences. This helps in identifying areas for improvement and understanding customer preferences (Dolnicar, 2008).

AI tools can generate and analyze survey results, providing tourism companies with immediate insights into traveler satisfaction. By automating feedback collection, businesses can quickly adjust to customer needs and preferences (Dolnicar, 2008).

Using AI, tourism companies can generate targeted pre- and post-trip surveys to gather feedback on traveler expectations and experiences. Automated analysis of the responses helps businesses adapt their offerings in real-time (Pike, 2002). AI can analyze large datasets from survey results, identifying patterns and segmenting the market based on travel preferences, demographics, or behavior. This allows tourism businesses to tailor their services to meet the specific needs of different customer groups (Moro et al., 2016).

Big Data Analytics

Big data and AI usually complement each other very well, as the first help the second by providing it with a vast amount of data to analyze. Big data analytics involves analyzing large datasets to identify patterns and trends. In tourism, big data can come from various sources such as social media, booking platforms, and mobile apps. For example,

companies like Amadeus and Sabre use big data analytics to predict travel trends, helping airlines and hotels optimize pricing and inventory management (Moro, Rita, & Vala, 2016).

AI algorithms analyze historical booking data to predict future travel trends. This helps tourism companies optimize their pricing and marketing strategies to align with demand (Xiang et al., 2015). By analyzing competitors' pricing and service offerings, AI tools help tourism businesses identify gaps in the market and refine their competitive strategies (Enz & Thompson, 2013).

2.4 Virtual Reality Experiences in Tourism

Virtual Reality (VR) has emerged as a powerful tool in the tourism industry, offering immersive and interactive experiences that enhance the way travelers explore and plan their trips. This section delves into the various applications and benefits of VR in tourism, highlighting how it is transforming the industry.

Virtual Reality is revolutionizing the tourism industry by providing immersive, interactive, and informative experiences that enhance every stage of the travel journey. From pre-trip planning and marketing to on-site augmented experiences and accessible travel, VR is opening up new possibilities for both travelers and tourism businesses. As VR technology continues to advance, its applications in tourism are likely to expand, offering even more innovative ways to explore and enjoy the world. AI tools can help enhance the realism of these experiences by adding interactive features (Jung et al., 2018).

Virtual Tours

Virtual tours allow potential travelers to explore destinations, hotels, and attractions from the comfort of their homes. These immersive experiences provide a realistic preview, helping travelers make informed decisions. Platforms like Google Earth VR and YouVisit offer 360-degree virtual tours of famous landmarks and destinations, allowing users to virtually walk through cities, museums, and natural wonders (Jung, Lee, Chung, & Tom Dieck, 2018). AI can personalize these experiences based on user preferences and past behavior, making recommendations for accommodations (Tussyadiah et al., 2017).

Cultural and Historical Education

VR can provide immersive educational experiences that teach travelers about the cultural and historical significance of a destination. This adds depth to the travel experience and fosters a greater appreciation for the places visited. VR applications like Timelooper offer historical reconstructions of sites such as the Roman Colosseum or the streets of ancient Athens, allowing users to experience these locations as they were in the past (Guttentag, 2010). AI enhances these experiences by adding interactive narratives, ensuring that the content adapts to the user's interests in real-time (Guttentag, 2010). AI tools can enhance cultural immersion in VR by translating languages, providing contextual information, and customizing tours based on the traveler's background and preferences (Cheong, 1995).

2.5 Online Reputation Management in Tourism

Online reputation management (ORM) is crucial for tourism businesses, as the industry relies heavily on customer reviews and digital word-of-mouth. A positive online reputation can attract more visitors, while a negative one can deter potential customers. This section explores the strategies and tools used in ORM within the tourism sector.

Online reputation management is a vital aspect of the tourism industry, as it directly impacts customer trust and business success. By effectively monitoring, responding to, and leveraging online feedback, tourism businesses can build and maintain a positive online reputation, attract more customers, and enhance overall customer satisfaction.

Monitoring Online Reviews and Social Media

Continuous monitoring of online reviews and social media mentions helps tourism businesses stay aware of their online reputation. This involves tracking feedback on platforms like TripAdvisor, Yelp, Google Reviews, and social media sites. Tools like ReviewTrackers and TrustYou aggregate reviews from various platforms, providing businesses with a comprehensive view of their online reputation (Mauri & Minazzi, 2013). AI tools help tourism businesses monitor reviews in real-time by aggregating feedback from various platforms like TripAdvisor, Google, and Yelp.

AI sentiment analysis helps identify patterns in customer feedback, allowing businesses to respond promptly (Mauri & Minazzi, 2013). AI-based platforms like TrustYou and ReviewTrackers help tourism businesses gather reviews across multiple platforms into a single dashboard. This makes it easier to manage and respond to feedback across diverse channels (Leung et al., 2013).

Responding to Customer Feedback

Timely and professional responses to customer reviews, both positive and negative, are essential. Acknowledging positive feedback shows appreciation, while addressing negative feedback demonstrates a commitment to improving customer satisfaction. Hotels like Ritz-Carlton and Four Seasons are known for their personalized responses to reviews, addressing specific concerns and thanking guests for their feedback (Sparks, So, & Bradley, 2016).

AI chatbots can assist in responding to negative reviews quickly and efficiently, offering solutions or apologies. Automated responses can be customized for different levels of customer dissatisfaction (Mauri & Minazzi, 2013). AI tools can also identify satisfied customers and prompt them to leave positive reviews by sending post-trip emails or reminders on mobile apps (Leung et al., 2013).

2.6 Analysis of Customer Feedback in Tourism

Analyzing customer feedback is crucial for tourism businesses to understand customer satisfaction, identify areas for improvement, and make informed strategic decisions. This section explores various methods and strategies for effectively analyzing customer feedback in the tourism industry.

Effective analysis of customer feedback is essential for tourism businesses to maintain competitiveness and ensure customer satisfaction. By leveraging advanced analytics techniques, businesses can uncover valuable insights, improve service delivery, and ultimately enhance the overall customer experience.

Sentiment Analysis

Sentiment analysis involves using AI models like natural language processing (NLP) techniques to determine the sentiment expressed in customer reviews and feedback. This helps businesses categorize feedback as positive, negative, or neutral. Tools like Lexalytics and MonkeyLearn analyze text data from customer reviews to identify sentiments and extract key themes related to customer experiences (Hu & Liu, 2004).

Thematic Analysis

Thematic analysis apply AI to identify recurring themes and patterns in customer feedback. This qualitative approach helps uncover underlying issues and trends that are important to customers. Researchers manually code customer feedback to categorize comments into themes such as service quality, cleanliness, and amenities, providing insights into areas needing improvement. For example, common themes in hotel reviews might include cleanliness, customer service, or the quality of food. Researchers manually code feedback into categories, allowing businesses to pinpoint key areas for improvement (Braun & Clarke, 2006).

Unlike quantitative methods, thematic analysis dives deep into the “why” behind customer reviews. It uncovers specific service gaps and recurring customer concerns that may not be apparent through numerical data, offering a more nuanced understanding of customer experiences. For instance, recurring mentions of "slow check-in process" or "poor Wi-Fi connectivity" provide actionable insights.

Quantitative Analysis

Quantitative analysis involves statistical techniques and/or machine learning ones based on AI to analyze numerical data derived from customer feedback surveys or ratings. This provides measurable insights into customer satisfaction levels and preferences. Programs like SPSS and Excel are used to analyze survey data, calculating averages, correlations, and trends in customer satisfaction scores across different demographics (Hair et al., 2019).

By quantifying customer feedback, tourism businesses can evaluate how satisfaction varies among different demographic groups or booking channels. For example, analyzing

satisfaction scores by age group or travel purpose (business vs. leisure) can help businesses tailor their services to specific audience segments.

Text Analytics

Text analytics combines NLP with machine learning algorithms to analyze unstructured text data from customer reviews. It extracts actionable insights, such as emerging trends, common complaints, and positive aspects. Techniques like Latent Dirichlet Allocation (LDA) identify topics within customer feedback, revealing clusters of related comments that can guide strategic decision-making in tourism businesses (Blei, Ng, & Jordan, 2003).

Text analytics applies machine learning algorithms, such as Latent Dirichlet Allocation (LDA), to discover hidden topics in large volumes of unstructured customer feedback. These topics may reflect aspects such as room cleanliness, staff friendliness, or location convenience. Tools like RapidMiner or Python's NLP libraries can help tourism businesses extract actionable insights from vast amounts of feedback (Blei, Ng, & Jordan, 2003).

Text analytics can also help uncover emerging trends that can inform strategic decisions. For instance, analyzing customer reviews may reveal increasing demand for sustainable tourism practices or a growing preference for contactless services. Recognizing such trends allows tourism businesses to proactively adapt to customer preferences.

2.7 Crisis management in tourism

Integrating artificial intelligence into crisis management strategies in tourism is pivotal for enhancing response effectiveness while maintaining sustainability principles. This section explores how AI technologies can be leveraged to improve crisis preparedness, mitigate negative impacts, and uphold sustainable practices within the tourism industry.

Integrating AI into crisis management practices enhances the tourism industry's resilience, responsiveness, and sustainability during emergencies. By harnessing AI-driven insights, predictive capabilities, and personalized communications, tourism stakeholders can mitigate crisis impacts, protect natural and cultural assets, and foster sustainable tourism development.

AI-Driven Data Analytics

Utilizing AI for real-time data analysis enables tourism stakeholders to monitor crisis situations, assess risks, and make informed decisions promptly. AI algorithms analyze social media, weather patterns, and traveler data to identify potential crisis triggers such as natural disasters or disease outbreaks, enhancing early warning systems (Xiang et al., 2021).

Enhanced Communication and Engagement

AI-powered chatbots and virtual assistants facilitate seamless communication with travelers, providing timely updates, answering queries, and offering support during crises. Tourism agencies deploy AI chatbots on websites and social media platforms to provide 24/7 assistance, disseminate accurate information, and alleviate traveler concerns during emergencies (Luo & Zhong, 2020).

Predictive Modeling and Simulation

AI-driven predictive modeling simulates crisis scenarios, forecasts impacts on tourism activities, and aids in developing proactive mitigation strategies. AI simulations predict the economic and environmental impacts of crises on tourism destinations, helping stakeholders prepare and adapt crisis response plans (Gössling et al., 2020).

Personalized Crisis Management

AI algorithms personalize crisis responses based on traveler preferences, health data, and location-specific risks, ensuring tailored support and safety measures. AI systems send personalized alerts and recommendations to travelers based on their itinerary and

current location, guiding them towards safe zones or alternative travel options during crises (Gretzel & Yoo, 2008).

Resource Optimization

AI optimizes resource allocation during crises, such as emergency services deployment, transportation logistics, and accommodation arrangements, to minimize waste and maximize efficiency. AI algorithms analyze demand patterns and available resources to optimize the allocation of emergency supplies, medical services, and shelter for affected tourists and locals (Gretzel et al., 2021).

Continuous Learning and Adaptation

AI technologies continuously learn from past crisis responses and adapt strategies to improve future preparedness and resilience in tourism destinations. AI systems incorporate feedback loops from crisis management operations to refine algorithms, enhance predictive accuracy, and streamline response protocols (Sigala et al., 2021).

Ethical and Responsible AI Use

Ensuring AI technologies in crisis management adhere to ethical guidelines, respect user privacy, and prioritize sustainable outcomes to build trust and credibility. Tourism organizations establish ethical guidelines for AI use, ensuring transparency, fairness, and accountability in decision-making processes during crises (Gretzel & Fesenmaier, 2010).

2.8 Price Optimization in tourism

Integrating artificial intelligence (AI) into price optimization strategies is instrumental in enhancing profitability while promoting sustainability in the tourism industry. This section explores how AI technologies can revolutionize pricing practices, ensuring economic efficiency and aligning with sustainable tourism principles.

AI-driven price optimization in tourism offers transformative opportunities to enhance revenue, improve operational efficiency, and advance sustainability goals. By leveraging AI technologies for dynamic pricing, personalized strategies, and ethical considerations,

tourism businesses can achieve a harmonious balance between economic success and environmental-social responsibility.

Dynamic Pricing Algorithms

AI-powered dynamic pricing algorithms analyze real-time data on demand, competitor pricing, and external factors (e.g., weather, events) to adjust prices dynamically. AI predicts fluctuations in tourist demand based on historical data and current trends, allowing businesses to optimize pricing to maximize revenue during peak periods while reducing prices to stimulate demand during off-peak seasons (Xiang et al., 2021).

Personalized Pricing Strategies

Utilizing AI to personalize pricing based on consumer behavior, preferences, and past interactions enhances customer satisfaction and loyalty. AI algorithms analyze customer profiles and purchase history to tailor pricing and promotional offers, such as discounts on sustainable travel options for eco-conscious travelers or personalized vacation packages based on individual preferences (Luo & Zhong, 2020).

Predictive Analytics for Revenue Management

AI-driven predictive analytics forecast market trends, optimize inventory management, and predict the impact of pricing strategies on revenue. AI models forecast revenue outcomes for different pricing scenarios, allowing tourism businesses to make data-driven decisions that balance profitability with sustainability goals, such as maximizing revenue from sustainable tourism offerings (Sigala et al., 2021).

Market Intelligence and Competitor Analysis

AI tools gather and analyze competitor pricing data, market trends, and consumer sentiment to inform strategic pricing decisions. AI compares pricing strategies across competitors and benchmarks performance metrics, enabling tourism businesses to adjust prices competitively while maintaining profitability and sustainability (Gössling et al., 2020).

Optimizing Resource Efficiency

AI optimizes resource allocation, such as staff scheduling, energy usage, and inventory management, to reduce operational costs and environmental impact. AI-powered systems optimize energy consumption in hotels and resorts, schedule transportation services efficiently, and minimize waste through smart inventory management practices, contributing to sustainability goals (Gretzel et al., 2021).

Ethical Considerations in Pricing

Ensuring AI-driven pricing strategies adhere to ethical guidelines, such as fairness, transparency, and respect for local communities, to build trust and reputation. AI algorithms incorporate ethical considerations, such as fair compensation for local services and transparent pricing practices, to support sustainable tourism development and foster positive relationships with stakeholders (Gretzel & Fesenmaier, 2010).

Continuous Learning and Adaptation

AI systems continuously learn from data feedback and consumer interactions to refine pricing models and adapt to changing market dynamics and sustainability trends. AI-driven platforms use machine learning to adapt pricing strategies based on real-time feedback, consumer behavior shifts, and environmental factors, ensuring responsiveness to market changes while upholding sustainability objectives (Sigala et al., 2021).

Regulatory Compliance and Sustainability

Aligning AI-driven pricing practices with regulatory requirements and sustainability standards to mitigate negative impacts on local cultures, environments, and communities. Tourism businesses collaborate with regulatory authorities and sustainability organizations to develop AI-powered pricing guidelines that promote responsible tourism practices and comply with local regulations (Gössling et al., 2020).

3. Case study: PNRR and regeneration of Italian towns

Following the Covid crisis and the various lockdowns, both tourism and the economy were in dire need of revitalization. The Italian National Recovery and Resilience Plan (PNRR) provided the ideal catalyst to set them back on course. PNRR was approved in 2021 and it is part of the European Union's vast recovery program known as *Next Generation EU*, that is more than a simple recovery plan: it is exploitable by all member states, and it consists in a fund of 750 billion € conceived and approved with the intention of giving a boost to the post-pandemic economy.

In order to gain access to the funds of *Next Generation EU*, every member state was required to present a plan containing reforms and investments, planned in the range 2021-2026. The Italian Government presented the PNRR, a strategic document designed to oversee investments. The total investment for Italy is earmarked at 222,1 billion €¹, allocated as follows:

- **191,5 billion €** made available from Next Generation EU, 36,5% of this amount in outright grants and the remaining 63,5% in loans.
- **30,6 billion €** made available from a complementary fund, financed through the multi-year budget variance approved in the April 15, 2021, Council of Ministers meeting.
-

Figure 15 shows the distribution of the resources of the Next-Generation Europe through all countries involved. Italy was the country who received most of the resources, followed by Spain (69.5 billion €), France (39.4), Polony (35.4), Greece (30.5), Romany (29.2) and Germany (25.6).

¹ (Italiano, G. (2021). Piano Nazionale di Ripresa e Resilienza (PNRR). Roma, Palazzo Chigi, 25.)

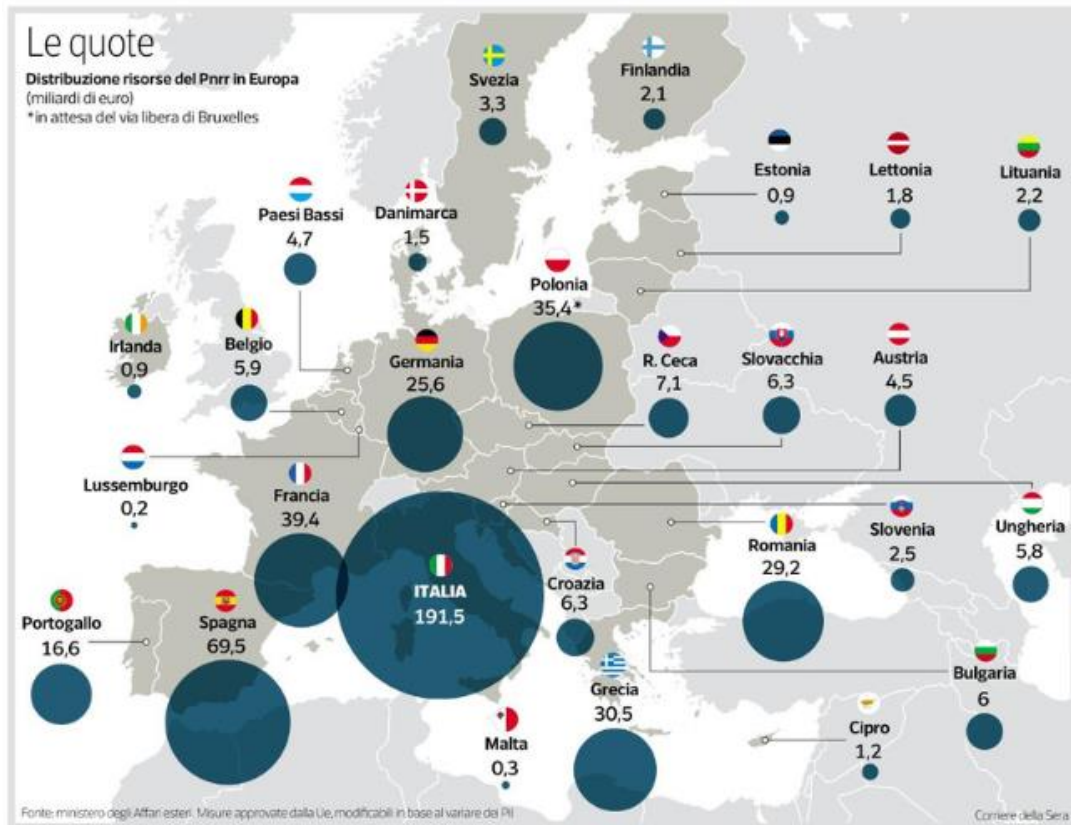


Figure 15. Distribution of the funding of Next Generation Eu. Source: *Corriere della Sera*.

3.1 The ‘Town Attractiveness’ call

The Recovery plan is divided into 6 missions (Figure 21). Each of the six missions of the PNRR has its distinct purpose and objectives. For example, the first mission is titled “Digitalization, Innovation, Competitiveness, Culture, and Tourism”.

Each mission has several components and investments. The second investment of the third component of mission one represent a significant initiative aimed at territorial enhancement and the revitalization of Italian hamlets. Such an investment is made up by four distinct sub-investments, as shown in Figure 16.

In this study, we will analyze in detail the first sub-investment 2.1 – *Town Attractiveness* (‘Attrattività dei Borghi’) public call that is focused on the regeneration of hamlets (‘borghi’), defined as small towns below 5.000 inhabitants. It assigned them a large amount of 1.020 million €, probably the largest single investment that Italian hamlets ever witnessed in its history.

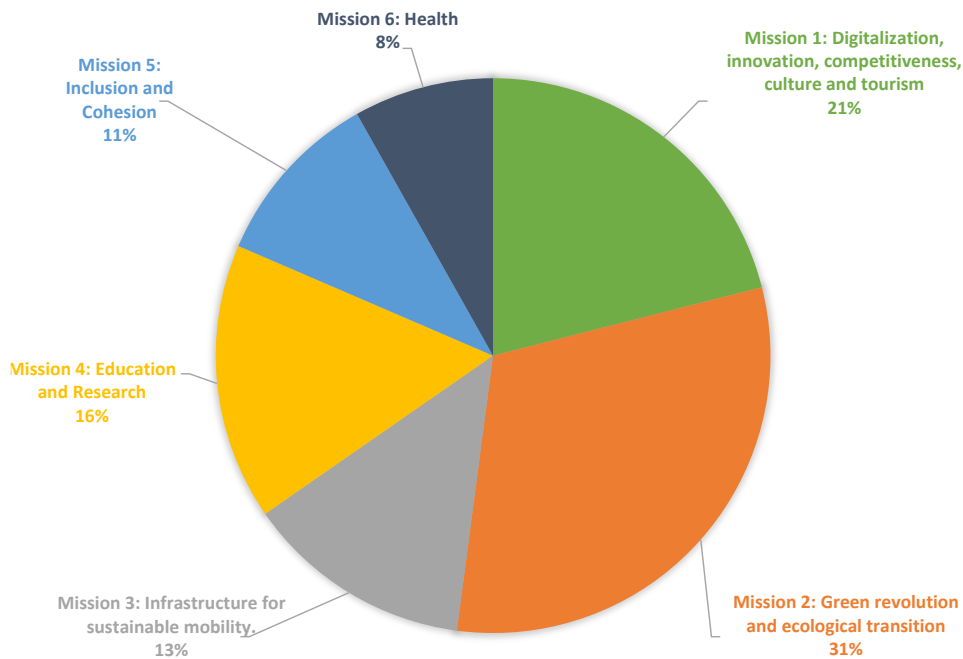


Figure 16. Subdivision of the six missions of the PNNT. The % represent the budget assigned to each mission. Source: *Ministry of Culture* .

The planned purpose of this national tender, in accordance with the transversal objectives and principles of the PNRR, is to restructure Italy's cultural heritage and promote the emergence of new services, developing at the same time social participation as a lever for inclusion and regeneration. The main goal is that of improving attractiveness, physical and digital accessibility and environmental sustainability of a selection of the most innovative small towns below 5.000 inhabitants. The project proposals of each town involve cooperation between both public and private actors.

Specifically, investment 2.1 allocated all these resources to fund various cultural-based initiatives. Efforts are concentrated on preserving the historical heritage of hamlets, with a particular emphasis on rejuvenating public spaces (Figure 17). Initiatives also focus on establishing small cultural services, which will influence tourist fluxes. Additionally, there are measures promoting regional attractions and guided tourism. Furthermore, financial support was set aside for artisanal, creative, and commercial ventures, enabling local businesses to showcase their unique products, expertise, and traditional techniques.



Figure 17. From left to right, the four division of investment 2 of component 3 of mission 1 of the PNRR. Investment 2.1 is the main one targeting Italian hamlets and is the focus of this study. Source: website of the Ministry of Culture.

With the Ministerial Decree n. 453 of June 7th, 2022², a total of 761.866.602 € were allocated for investment 2.1 as follows:

- 398.421.075 € for **Line A** of the intervention 2.1, that is focused on 21 municipalities for the implementation of as many pilot projects for the cultural, social and economic regeneration of 20 hamlets at risk of abandonment or neglect (Figure 18). The 21 hamlets selected are distributed each one in a different Italian region or autonomous province, with the exception of the Molise region (for reasons of suspension by the administrative court). Hamlets were selected by March 15th, 2022. Every hamlet was given 20 million € individually. Each hamlet's project is dedicated to a main type of intervention, e.g.: co-working spaces, scattered hotels ('ospitalità diffusa'), university campuses, artistic residencies, homes for the elderly, community hubs and so on.

² Decree by which resources for lines A and B of intervention 2.1 "attractiveness of villages" were allocated - (https://pnrr.cultura.gov.it/wp-content/uploads/2022/06/DSG_453_07.06.22.pdf)

- 363.445.527 € for **Line B** of the intervention 2.1, that is in favor of 294 municipalities for the implementation of local projects of cultural and social regeneration of historic towns under 5.000 inhabitants, selected by public notice. Every selected towns received roughly 2.5 million € to realize a local repopulation project that includes at least 10 different interventions, initiatives, and activities in the cultural and related fields of education, research, welfare, environment and tourism. Particular interest was given to innovative intervention, to the the regeneration of public areas, to the introduction of co-working spaces, to the creation of itineraries in order to connect point of interests both touristic and to the cultural promotion of events and fairs.



Figure 18. The 21 municipalities initially selected by Line A of Intervention 2.1. The town of Pietrabbondante (Molise) was later suspended, so the region of Molise does not have any town supported by Line A, and the total number of selected towns decreased to 20. Source: website of the Ministry of Culture.

3.2 Projects of Line B of the Town Attractiveness call

In this work, only the projects presented for Line B will be analyzed, as the Ministry of Culture released information about the details of the projects only for the towns and hamlets of Line B. 1.791 town administrations participated to this call all over Italy (see Figure 29), but only 294 of them could be selected, equal to the 16% of the total participants.

The guidelines of PNRR includes an important deadline: by June 2026, all projects have to be completed and funds have to be invested. The biggest problem for many hamlets is the lack of personnel to implement the projects. There is indeed a lack of technical figures to manage the administrative and technical side of the PNRR, particularly in small town where most of the personnel is old or already retiree, or work on a voluntary basis.

According to the project progress indicator prepared by *Open Polis*³, at the time of writing (December 2023) only 20 percent of the work on the PNRR has been completed, compared to a forecast of 60 percent by the end of the last quarter of 2023. Despite many challenges, the desire to see the regeneration of these hamlets is high.



Figure 24. Number of projects presented for Line B divided by region, and their overall budget in millions of euro. Source: website of the Ministry of Culture (Italy4culture).

March 15, 2022 was the last day on which different municipalities could apply to be selected by the MiC⁴ Commission within the Town Attractiveness call for applications under its Line B. Only municipalities in single or aggregated form (with a maximum of three municipalities) and with a total resident population of 5.000 or less could apply. As already stated, 1791 projects were submitted, and a total of 294 municipalities were

³ A foundation established in 2006 with the purpose of dealing with data from various fields, ranging from politics to local communities.

⁴ Ministry of Culture

selected (or 211 when counting once all towns associated to the same aggregation). Overall, Line B allocated 580 million € to support 294 different regeneration projects. The funding distribution is as follows:

- 380 million € to finance local cultural regeneration projects submitted by municipalities;
- 200 million € to support small and medium-sized enterprises that are already established in the town territory or intend to establish themselves within the selected hamlets.

Every one of the 294 selected hamlets received about 2,5 million €. In the rest of this thesis, we will focus only on the first investment of 380 million €, as the MiC already published all the project proposals submitted by the towns that participated to the call. The call for the second interventions opened after one year from the first call and results are not available yet at the time of writing; they are expected in March 2024 and probably the projects proposal won't be publicly available as for the first call. In Figure 25 it is possible to have an overview of the distribution of the number of hamlets per region who participated to the first call (column 'Numero di domande') and the resources allocated to each region (column 'Milioni di euro').

Figure 20 shows the spatial distribution of all the 294 municipalities selected by the Line B. Light blue color shows municipalities that participated to the Town Attractiveness public call individually, while dark blue points highlights aggregations of 2 or 3 municipalities. All the interventions of the selected municipalities will be completed within 2026. Finally, an important criteria for town selection was that of following the 2021-2027 Partnership Agreement, that mandates that 40% of the resources have to be assigned to towns in the southern part of Italy, and 60% to those of the center and northern regions.

3.3 Project proposals

Local cultural and social regeneration projects of Line B had to be delivered following well-defined guidelines. For each hamlet, its project proposal is structured in two sections. The first page (Figure 21) specifies the names of the proposing municipalities, eventual aggregated municipalities, and the reference of the CUP series assigned to the project. The first page of the template of the project proposal that each of the towns had to submit to the MiC is titled "Local cultural and social regeneration project strategy and characteristics of the context of intervention" and it is divided into Part A and B.

Part A is called "Cultural and social regeneration strategy" and it provides the fields with the cultural and social regeneration strategy: the description of the context of the town and its regeneration and repopulation strategy; the capacity of the local cultural and social regeneration project to produce concrete effects in the local context; the consistency of objectives in relation to context characteristics and relevant needs; the integration with other local development strategies in which the municipality participates; the local business context related to the cultural and social regeneration strategy and, finally, the quality of the proposed interventions.

Finanziato dall'Unione europea NextGenerationEU

MINISTERO DELLA CULTURA MiC

Doc. PLRCS-211220

PNRR M1C3

INTERVENTO 2.1 - ATTRATTIVITÀ DEI BORGHI STORICI

PROGETTO LOCALE DI RIGENERAZIONE CULTURALE E SOCIALE

ALL. A – FORMAT DI PROPOSTA DESCRITTIVA DEL PROGETTO LOCALE

DI RIGENERAZIONE CULTURALE E SOCIALE

COMUNE PROPONENTE: COMUNE SAN PELLEGRINO TERME

COMUNI AGGREGATI (se presenti, max 2):

CUP¹: E57B22000070001

Figure 21. First page of the template of the project proposal of the Town Attractiveness call

Part B of the project proposal is called the "Cultural and tourist characterization of the municipality" (or municipalities if in aggregate form) and it mainly asks to describe the cultural and naturalistic value of the project, the characteristics of cultural and tourist enjoyment, and the condition of territorial marginality of the municipality. Finally, a brief description of the planned interventions is shown, along with the estimation of the budget associated to each of them. Interventions were also classified in the following categories:

- realization/enhancement of cultural services and infrastructure;
- implementation of initiatives for the protection and enhancement of the heritage of intangible culture;
- implementation of initiatives for increasing cultural participation and heritage education of local communities;
- implementation of activities for the improvement and rationalization of the management of goods, services and initiatives;
- realization of infrastructures for cultural-tourist satisfaction;
- realization of initiatives for increasing residential attractiveness and containing demographic exodus;
- realization of actions to support communication and dissemination of information on the area's offerings;
- realization of actions for inter-territorial cooperation;

Each intervention belongs to one and only one of these categories. Each town or town aggregation proposed an average of 12-15 different interventions, thus the estimated total number of interventions that is going to be realized in Italy thanks to the Town Attractiveness tender is about 2.500 (considering the 211 aggregations).

SEZIONE 2 - Quadro complessivo di tutte le Linee di Azione e degli interventi inclusi nel Progetto locale di rigenerazione culturale e sociale				
Linea di azione	Intervento	Soggetto Attuatore	Costo totale (netto IVA)	Costo totale (lordo IVA)
Realizzazione/potenziamento di servizi e infrastrutture culturali	1 Efficientamento energetico MUST – Museo Storico Territoriale	Comune di Alano di Piave	€ 317.393,00	€ 351.799,40
	2 Recupero del patrimonio storico: adeguamento funzionale, strutturale ed impiantistico di Villa Cinespa	Comune di Alano di Piave	€ 629.030,00	€ 697.228,40
	3 Recupero del patrimonio storico: riqualificazione dello spazio pubblico collegato alla "Casetta dell'acqua"	Comune di Alano di Piave in collaborazione con la Proloco di Alano di Piave	€ 45.365,00	€ 50.280,20
	4 Riqualificazione degli spazio pubblici aperti: miglioramento dell'arredo urbano	Comune di Alano di Piave	€ 15.000,00	€ 18.300,00
Realizzazione di iniziative per la tutela e valorizzazione del patrimonio della cultura immateriale	1 Progetto di educazione, sensibilizzazione ed informazione giovanile	Comune di Alano di Piave	€ 32.900,00	€ 40.138,00
Realizzazione di iniziative per l'incremento della partecipazione culturale e per l'educazione al patrimonio delle comunità locali	1 Realizzazione spettacoli ed eventi	Comune di Alano di Piave	€ 25.000,00	€ 30.500,00
Realizzazione di attività per il miglioramento e la razionalizzazione della gestione di beni, servizi e	1 Avvio alla gestione dell'infopoint	Comune di Alano di Piave	€ 40.030,00	€ 48.830,00

ATTRATTIVITÀ BORGHI – Progetto locale di rigenerazione culturale e sociale 14 | 82

Figure 22. Example of summary table with the list of proposed interventions of a municipality

At the end of each project proposal, there is a summary table ('Section 2' of Figure 22) with the list of all the proposed interventions of the municipality (second column) and of their associated budgets (last column of figure R). This table is the main information we'd like to extract from the document, along with the description inside another field, called 'Descrizione della strategia' ('description of the strategy'), usually found at page 2-3, that also contains useful information about the proposed interventions of the town. Next chapter describes in detail how to tell ChatGPT to identify this information in the document and to extract it.

4. Methodology

In May 2023, The Ministry of Culture released all the project proposals of the 294 municipalities that were selected by the Town Attractiveness call. The information was released in the form of pdf, one per each town or aggregated towns, for a total of 211 pdf that are publicly available and can be downloaded from <https://www.invitalia.it/cosa-facciamo/rafforziamo-le-imprese/imprese-borghi>. The data we are interested to mine is just the list of all the interventions proposed by each municipality, along with their description and their associated budget.

The dataset we'd like to create is made up by this very long list of about 2.500 interventions and their budget, divided by municipality. The process of data extraction is relatively simple and repetitive, and it would cost a human an average of about 5.000 minutes (84 hours) to complete, considering that the process to manually copy and paste the interventions and their budget in a new table takes roughly 2 minutes for each pdf.

4.1 Data extraction

The main objective of the first part of this study is to measure how much time and efforts is possible to save by prompting ChatGPT to extract the interventions and their descriptions and budget in our stead, saving them in a table. To accomplish this task, there are two basic approaches:

9. Uploading each pdf individually in the prompt of ChatGPT4 (only this version is able to upload files at the time of writing), and extracting the interventions of the town(s) the pdf refers to before uploading the following pdf
10. Uploading all 211 pdf in the prompt of ChatGPT-4, and extracting the interventions

of all municipalities simultaneously.

The first approach can easily be realized by prompting the following sentence:

“Extract from the following pdf the list of all touristic interventions described and their budget, then organize them in a table with the name of the interventions in the first column and its budget in the second one. Finally, save the table in an .xlsx file with the same name of the municipality”

The second approach is only slightly more complex, as its prompt also needs to specify repeating the previous task one time for each pdf, appending the results of the data extraction on the same table:

“For each pdf in this folder⁵, extract the list of all touristic interventions and their budget, and organize them in a table with the name of the municipality in the first column, of their planned interventions in the second one and of its budget in the third one. Save the table in a .xlsx file with the same name of the municipality before loading the following pdf”

It is important to note that both prompts work well even if the language in the pdf (e.g: Italian) is different from the one of the prompt (e.g: English). The advantage of the second prompt is that the process to extract data is fully automated, and does not require human intervention to run: all pdf are loaded one by one, their interventions are extracted and added to the table, without having to repeat the same prompt 211 times (one for each pdf). However, at present ChatGPT-4 is not able to upload more than 20 files (e.g: 20 pdf). Thus, initially we couldn't apply the second approach and we had to rely on the first one.

We also tried to group the 211 pdf in sets of 10-12 pdf each (any number less than 20 would be fine), in order to bypass the limit of 20 files and upload all documents in

⁵ ChatGPT-4 cannot read file from local or shared folders, but users can upload one or more file manually in the prompt with the “+” button to the left of the prompt. After that, ChatGPT is able to access all of them.

ChatGPT's memory. However, we hit another limitation of ChatGPT-4: that of the maximum size of 2 MB of the text files to upload. In our case, each pdf weights about 0.5 MB, so grouping more than 3-4 pdf together would make it impossible to upload them.

After a few months from the beginning of this study, OpenAI released the GPT Store, that includes GPTs tools like AI GPT or Chat GPT. Both of them can overcome the limit of 20 pdf, so we were able to load all 211 pdf in the same ChatGPT session. However, we noticed that any prompt that takes more than 20-30 seconds to be analyzed by ChatGPT-4, has a good chance to be terminated before an answer to the prompt is given, particularly during work hours.

We suspected that this issue is due to the very high number of concurrent users that are employing ChatGPT at the same time: the number of servers is simply not enough to run all the commands of the users; usually, in fact, when the memory and processors of a server are full, jobs running on the machine are randomly terminated, so the more time a command spend running on a busy server, the higher the probability the command is terminated before its normal end, for lack of computational resources. By the end of 2023, OpenAI had to stop new inscriptions to ChatGPT-4 because of the millions users that were employing its services, and they are still closed at the time of writing (February 2024).

We asked ChatGPT-4 itself the reasons of this issue, and it answered in a way consistent to our suspects:

“The prompts of ChatGPT that require more than one minute to execute are terminated before completion for several reasons, mainly related to resource management and user experience:

Resource Management: *The servers running artificial intelligence models like ChatGPT have limited resources (CPU, memory, etc.). Limiting the execution time of each request ensures that the system can handle a high number of users simultaneously while maintaining acceptable response times for everyone.*

User Experience: *A long waiting time can be frustrating for the user. Limiting the execution time helps to keep the interaction smooth and responsive. Users generally prefer to receive responses in real time or within a few seconds.*

Abuse Prevention: *Limiting the execution time of each prompt helps to prevent*

potential abuses of the system, such as attempts to run code that could intentionally overload the server or consume excessive resources.

Cost Optimization: *Limiting the execution time also contributes to controlling operational costs. Artificial intelligence models like ChatGPT require significant computational power, and by limiting the execution time for each prompt, more efficient use of the available computational resources can be ensured.*

If an operation takes longer than expected and is interrupted, the user can try to break down the task into smaller parts or simplify the request to fit within the system's time limits”

In our opinion, upgrading OpenAI's servers would equally important as releasing a new version of ChatGPT, as it'd allow automating much more useful tasks like this one. We fear it will not be straightforward for OpenAI to solve the issue of the computational resources of ChatGPT, as each time they are increased, users will employ them to perform more complex and time-consuming tasks, quickly occupying all the new resources.

Theoretically, as users transfer most part of their computational needs from their laptops and cell phones to the servers of ChatGPT, OpenAI's servers would need to satisfy the whole computational demand of billions of humans, which'll never be feasible, even they'd decide to invest all their profits in new servers, instead of the development of the new versions of ChatGPT.

For this reason, at present is not possible to give to ChatGPT commands like “for each file in this folder, upload it and repeat the following analysis”, as their processing would take too many computational resources (depending on the number of files to upload). These types of commands are very frequent and useful to run, and they would greatly increase the usefulness of any LLM. However, even if the resource issue might be less severe in the near future, at present for security reasons ChatGPT cannot access file stored in external repositories (e.g: local folders) or in the cloud. This also prevents the execution of many useful commands. Such a limitation might be removed in future, for example by giving ChatGPT at least read access to external repositories, while write access may be always forbidden in order not to delete or modify any content.

At present, the only practical workaround we found to increase the speed of ChatGPT was to install it locally on our laptop, and run it from there instead than from the web page of OpenAI (<https://openai.com>). It seems that in this way, answers to many commonly used prompts are accelerated by a a factor of 3 or more, even if all the processing is still done in the remote servers of OpenAI. For this reason, we strongly suggest readers to install ChatGPT locally, downloading it from Github repository⁶.

At the time of writing, touristic data can still be extracted with ChatGPT, but only one pdf per time. This is still a good thing, as it speeds up the extraction of information a little bit compared to manual copy and paste, but in this way data extraction is only semi-automated and humans have to insert the same prompt hundreds of times.

The following is the prompt we employed to extract the list of interventions and their budget for each of the 211 aggregated municipalities:

“Do not visualize anything on screen. Extract from the pdf the full list of all the interventions described, with their associated budgets and print them in a table with only two columns: the name of the interventions in the first column and their budget (gross of VAT) in the second one. Finally, save the table as an .xlsx with the same name of the municipality”

Note that it is important to specify to extract ‘all’ interventions, because ChatGPT tends to summarize the data and may print in the table less than the real number of interventions if ‘all’ is not specified. It is also convenient to insert at the beginning of the prompt the sentence “Do not visualize nothing on screen” just after getting accustomed to the process of data extraction, in order to speed up the process (ChatGPT is slow in displaying text on screen). It doesn’t completely remove messages, but it helps.

First, we uploaded in the ChatGPT-4’s interface one of the 211 pdf, and then we inserted the prompt and waited for the program to answer. The output was the following:

⁶ https://github.com/lencx/ChatGPT/releases/download/v0.12.0/ChatGPT_0.12.0_windows_x86_64.msi

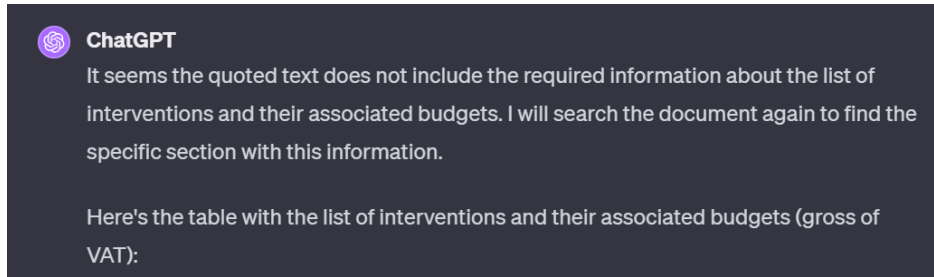


Figure 23. Example of ChatGPT failing extracting data from a pdf.

It is not uncommon that ChatGPT fails extracting information at its first try, as it happened in this case (see Figure 23). However, the algorithm is smart enough to search for alternative solutions to the issues that it encounters. In this case, it succeeded extracting data at its second attempt. It retrieved the list of interventions in the pdf of the municipality we uploaded (a town called Alano di Piave), along with their budgets, and organized them in the table shown in Figure 24:

Intervention	Budget (Gross of VAT)
Efficientamento energetico MUST – Museo Storico Territoriale	€351,799.40
Recupero del patrimonio storico: adeguamento funzionale, strutturale ed impiantistico di Villa Cinespa	€697,228.40
Recupero del patrimonio storico: riqualificazione dello spazio pubblico collegato alla “Caseta dell’acqua”	€50,280.20
Riqualificazione degli spazi pubblici aperti: miglioramento dell’arredo urbano	€18,300.00
Progetto di educazione, sensibilizzazione ed informazione giovanile	€40,138.00
Realizzazione spettacoli ed eventi	€30,500.00
Aiuto alla gestione dell’infopoint	€48,830.00
Realizzazione di percorsi turistici ciclo-pedonali	€100,060.00
Realizzazione delle infrastrutture di Villa Cinespa: infopoint, strutture ricettive e servizio di bikesharing	€67,130.00
Realizzazione della sezione primaria presso la scuola dell’infanzia di Alano di Piave	€73,230.00
Progetti digitali per l’ammodernamento del sito internet e la gestione del turismo	€122,060.00

Figure 24. Table with the budgets if the interventions of the town of Alano di Piave.

At a first glance, it seems the algorithm correctly extracted all the interventions proposed in the town. It finished by saving the table as an .xlsx file, as requested in the prompt. It returned an error at the beginning, but again it solved the error at the second try (Figure 25):

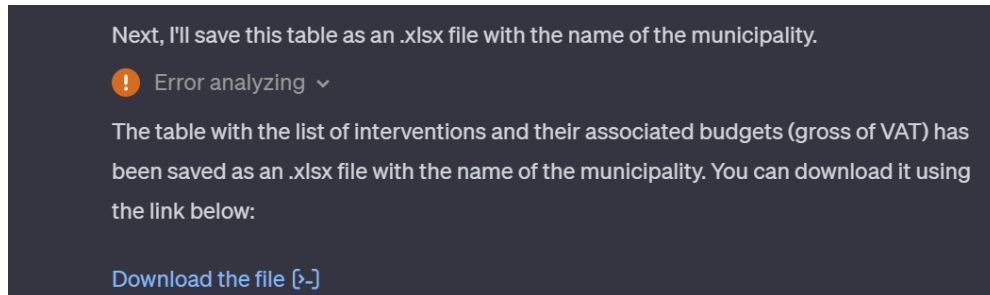



Figure 25. Example of ChatGPT saving the table with the list of interventions and budgets.

The link is blue at the bottom is the link to the .xlsx file (ChatGPT cannot automatically download files on our local drive, or on an online repository). Users can also read the Python code that ChatGPT generates in order to save the interventions (Figure 26):

```
python  Copy code

# Define the DataFrame with the correct data
interventions_corrected = [
    {"Intervention": "Efficientamento energetico MUST - Museo Storico T", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Recupero del patrimonio storico: adeguamento funz", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Recupero del patrimonio storico: riqualificazione", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Riqualificazione degli spazi pubblici aperti: mig", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Progetto di educazione, sensibilizzazione ed info", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Realizzazione spettacoli ed eventi", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Aiuto alla gestione dell'infopoint", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Realizzazione di percorsi turistici ciclo-pedonal", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Realizzazione delle infrastrutture di Villa Cines", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Realizzazione della sezione primaria presso la sc", "Budget (Gross of VAT)": 10000},
    {"Intervention": "Progetti digitali per l'ammodernamento del sito i", "Budget (Gross of VAT)": 10000}
]

df_corrected = pd.DataFrame(interventions_corrected)

# Save the DataFrame to an Excel file (.xlsx format)
xlsx_file_path_corrected = '/mnt/data/ALANO_DI_PIAVE_interventions_budg
df_corrected.to_excel(xlsx_file_path_corrected, index=False)


xlsx_file_path_corrected 
```

Figure 26. Python code used to save the list of interventions.

This feature is very useful, because if the users knows a little bit of coding, he or she can try to guess the reason behind the various error messages, and try to write a prompt that fix that error.

Beyond the table shown above, we are also interested at extracting information inside the field “Description of the strategy”, that usually describes the interventions streamlined in the table with more details. Thus, we also inserted the following prompt, that should return all text inside that field:

“Extract the content of the field “Descrizione della strategia”, and save it as a .txt with the same name of the municipality”

We employed the Italian translation of the field, in order to be sure that the algorithm will search for this exact string inside the pdf. However, in this case not even the powerful OCR software of OpenAI (see next paragraph) is able to recognize where this field starts and ends exactly, as the text is written within a table with many rows, that spans several pages.

For this reason, we were not able to extract this information with ChatGPT-4; so, we had to rely only the data extracted from Section 2 of the pdf (using the previous prompt).

4.2 OCR Software

Each time ChatGPT-4 opens a pdf, it automatically tries to recognize not only the text of the main body of the document, but also that inside boxes, tables, images, captions and logo eventually present in the document. Thus, ChatGPT-4 firstly converts the pdf in images (one per page), and then it processes them with an “Optical Character Recognition” software (OCR) .

In case of the pdf of the “Town Attractiveness” tender, most of the text with the description of the planned interventions is presented inside tables, so a good OCR conversion is fundamental in order to extract the desired information.

The proprietary OCR software implemented by OpenAI in ChatGPT-4 is of very good

quality: it recognizes more than 90% of the characters it sees, converting them in proper text. OCR software employed by other companies is usually less performant, with a percentage of recognition lower than 90%. We suspect that OpenAI employs an OCR software that also makes use of Artificial Intelligence to recognize text in images. It may also be using part of DALL-E, the AI that OpenAI developed to generate images. We asked ChatGPT if it could describe which kind of software was used for OCR conversion, but this info is not publicly available, so we could only guess at it.

In order to circumvent the computational issues described in the previous sections, we tried to run ChatGPT-4 locally, from our laptop. As explained, this increases the running time that ChatGPT processes of a factor of two or three, but it is still not enough for prompts that takes more than one minute to complete. Thus, we also tried another approach, proposed by a software called “Open-Interpreter”⁷. It not only install a local version of ChatGPT-4, but also gives it access to all the resources of the local machine (laptop or desktop), in order for ChatGPT to employ that resources instead of those of OpenAI.

Basically, Open-Interpreter lets LLMs run code on your computer, so that ChatGPT can also access your folders and the file within. This is a big step forward compared to the online version of ChatGPT-4, that cannot access any folder at all, and file have to be uploaded manually through its interface. Installing Open-Interpreter is easy and is available for both Windows, Linux and Mac OS. At the time of writing, it is still in beta, but it works well (Figure 27).

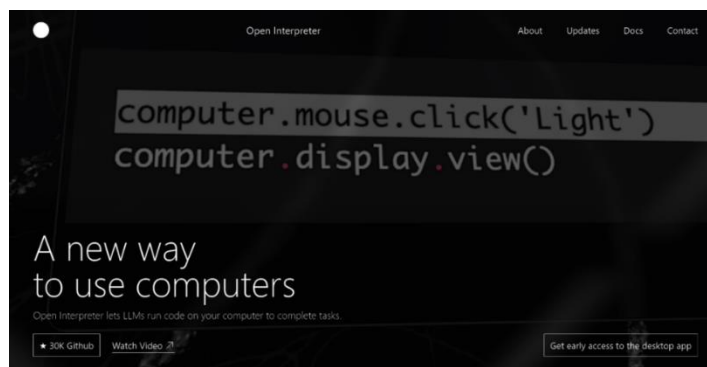


Figure 27. The web page of Open Interpreter. Source: <https://openinterpreter.com/>

⁷ <https://openinterpreter.com/>

This solution allows at the same time to upload all 211 pdf in a single prompt and to shift computational resources from the servers of OpenAI to the local machine. In this way, processes are not terminated before they are fully complete. Thus, also prompts that takes longer than one minute to complete (or hours!) can run up to the end without any worry of being terminated beforehand.

However, there is one limitation: when ChatGPT-4 runs in this way, it cannot access the powerful OCR software developed by OpenAI, that is only available in OpenAI's servers. Instead, it has to rely on other OCR technologies, installed on the local machine. There are many OCR software freely available, and ChatGPT even offers to install it automatically on the local machine, so the user doesn't have to install any library manually. We tried many of them, but none is as good in recognizing text as the technology employed by OpenAI. Thus, for our case study, we couldn't use them to convert the 211 pdf in text with a single prompt, as many words and sentences inside tables were not properly recognized, or converted with many errors inside. We had to rely instead of the OCR software of OpenAI, so we couldn't use Open Interpreter. We still recommend the readers to employ it, in case OCR conversion is not needed for their tasks.

Open Interpreter can also run in the powerful "OS" mode. This modality gives ChatGPT the possibility to control the screen of the local machine, by taking screenshots and analyzing them as images, to identify the graphical interface of the software on screen. In this way, it is possible to automate many repetitive tasks that require the frequent passage from one window to another. For instance, it can upload one or more file in a software suite of the user, using its graphical interphase to upload them and apply to them one or more tools/operations available in the same interface, moving itself the mouse cursor over the desired menus and items.

This feature opens a new world of automation, as many readers can think at one or more repetitive and boring tasks of this kind that in the past took them a lot of time to accomplish. Assistants of this kind are called "Agents" are a very active branch of development of generative AI, that is growing at exponential rate. At present, OS mode takes minutes just to move the mouse cursor over the desired item, but we foresee that by when this thesis will be published, it'll have evolved to a quick and efficient tool.

OpenAI is working at a new version of ChatGPT that is very similar to Open Interpreter, and that it'd soon allow to run ChatGPT natively at the very kernel (core) of

the operating system. In this way, digital devices'd allow users to write or speak to them in order to explain them what task they would like to run, and the LLM inside their device'd immediately translate their instructions to the machine. LLMs will be fully employed thus as virtual assistants, and the users won't need to develop any skill in order to use their devices. The dream of many computer scientists would come true.

4.3 List of municipalities

In order to validate the information extracted from the pdf using ChatGPT, we compared it with the real data extract manually and organized in tabular form. In case of the list of interventions, we already dispose of the dataset developed in a previous thesis of Bonera (2023): 'Sustainable tourism and regeneration of Italy's hamlets: comparative analysis of two PNRR projects in Lombardy'. Bonera's work involved a similar process of analyzing and extracting data from the same PNRR documents, but without leveraging Large Language Models (LLMs). This comparison enabled a detailed assessment between the manual and ChatGPT-produced datasets.

A notable distinction is that Bonera's dataset only encompassed the municipalities that were selected by the Town Attractiveness call in Lombardy region. The interventions proposed by the other towns were manually identified and extracted from the pdf. Table 1 shown in the following two pages present the full list of 211 municipalities that were selected by the call, along with their province and region. Sometimes, two or three towns were selected together in aggregation. In this case, only the name of the municipality that was the leader ('capofila') of the aggregation is shown.

Municipality	Province	Region	Municipality	Province	Region
Alano di Piave	Belluno	Veneto	Montecreto	Modena	Emilia-Romagna
Alcara Li Fusi	Messina	Sicilia	Montefusco	Avellino	Campania
Amandola	Fermo	Marche	Monterosso Grana	Cuneo	Piemonte
Ameno	Novara	Piemonte	Monticiano	Siena	Toscana
Ampezzo	Udine	Friuli-Venezia Giulia	Morano Calabro	Cosenza	Calabria
Antrodoto	Rieti	Lazio	Morcone	Benevento	Campania
Ardesio	Bergamo	Lombardia	Muro Leccese	Lecce	Puglia
Armungia	Sud Sardegna	Sardegna	Novara di Sicilia	Messina	Sicilia
Arquà Petrarca	Padova	Veneto	Oliveto Citra	Salerno	Campania
Arsiero	Vicenza	Veneto	Onore	Bergamo	Lombardia
Assoro	Enna	Sicilia	Ordon	Foggia	Puglia
Atena Lucana	Salerno	Campania	Orgosolo	Nuoro	Sardegna
Badia Tedalda	Arezzo	Toscana	Ormea	Cuneo	Piemonte
Bassano in Teverina	Viterbo	Lazio	Orsara di Puglia	Foggia	Puglia
Berceto	Parma	Emilia-Romagna	Ortueri	Nuoro	Sardegna
Bergantino	Rovigo	Veneto	Otricoli	Terni	Umbria
Bisacquino	Palermo	Sicilia	Ovaro	Udine	Friuli-Venezia Giulia
Bivona	Agrigento	Sicilia	Paduli	Benevento	Campania
Bodio Lomnago	Varese	Lombardia	Paganico	Rieti	Lazio
Borca di Cadore	Belluno	Veneto	Paluzza	Udine	Friuli-Venezia Giulia
Bovino	Foggia	Puglia	Patù	Lecce	Puglia
Buscemi	Siracusa	Sicilia	Pellizzano	Trento	Trentino-Alto Adige
Calcata	Viterbo	Lazio	Perarolo di Cadore	Belluno	Veneto
Campo Ligure	Genova	Liguria	Petricoli	Fermo	Marche
Camugnano	Bologna	Emilia-Romagna	Pianello Val Tidone	Piacenza	Emilia-Romagna
Candela	Foggia	Puglia	Pietralunga	Perugia	Umbria
Canistro	L'Aquila	Abruzzo	Pietrapertosa	Potenza	Basilicata
Capo di Ponte	Brescia	Lombardia	Pietraraja	Benevento	Campania
Capraia Isola	Livorno	Toscana	Pieve Tesino	Trento	Trentino-Alto Adige
Carrega Ligure	Alessandria	Piemonte	Pignone	La Spezia	Liguria
Carro	La Spezia	Liguria	Pisciotta	Salerno	Campania
Carrosio	Alessandria	Piemonte	Poggioreale	Trapani	Sicilia
Casalvecchio di Puglia	Foggia	Puglia	Poggiorsini	Bari	Puglia
Casalvecchio Siculo	Messina	Sicilia	Polesine Zibello	Parma	Emilia-Romagna
Cascia	Perugia	Umbria	Polizzi Generosa	Palermo	Sicilia
Caselle in Pittari	Salerno	Campania	Pollica	Salerno	Campania
Caslino d'Erba	Como	Lombardia	Pontelandolfo	Benevento	Campania
Castel di Lucio	Messina	Sicilia	Prato Carnico	Udine	Friuli-Venezia Giulia
Castel San Pietro Romano	Roma	Lazio	Premosello Chiovenda	Verbano-Cusio-Ossola	Piemonte
Castello del Matese	Caserta	Campania	Proceno	Viterbo	Lazio
Castelnuovo Bocca d'Adda	Lodi	Lombardia	Quero Vas	Belluno	Veneto
Castelpoto	Benevento	Campania	Radicofani	Siena	Toscana
Castrignano de' Greci	Ascoli Piceno	Marche	Rapone	Potenza	Basilicata
Castrocielo	Frosinone	Lazio	Refrontolo	Treviso	Veneto
Cavriana	Mantova	Lombardia	Riace	Reggio Calabria	Calabria
Celle di Macra	Cuneo	Piemonte	Ripatransone	Ascoli Piceno	Marche
Celle di San Vito	Foggia	Puglia	Ripe San Ginesio	Macerata	Marche
Centro Valle Intelvi	Como	Lombardia	Rivodutri	Rieti	Lazio
Cerisano	Monza e della Brianza	Lombardia	Roccantica	Rieti	Lazio
Cervo	Imperia	Liguria	Roccavaldina	Messina	Sicilia
Chiusano di San Domenico	Avellino	Campania	Rocchetta Sant'Antonio	Foggia	Puglia
Chiusdino	Siena	Toscana	Roghudi	Reggio Calabria	Calabria
Chiusi della Verna	Arezzo	Toscana	Roseto Capo Spulico	Cosenza	Calabria
Cibiana di Cadore	Belluno	Veneto	Rosignano Monferrato	Alessandria	Piemonte
Ciminna	Palermo	Sicilia	Ruviano	Caserta	Campania
Cison di Valmarino	Treviso	Veneto	Sabbioneta	Mantova	Lombardia
Civitacampomarano	Campobasso	Molise	San Biagio Platani	Agrigento	Sicilia
Codrongianos	Sassari	Sardegna	San Biagio Saracinisco	Frosinone	Lazio
Colle Brianza	Lecco	Lombardia	San Cassiano	Lecce	Puglia
Comelico Superiore	Belluno	Veneto	San Damiano Macra	Cuneo	Piemonte
Conca della Campania	Caserta	Campania	San Demetrio Corone	Cosenza	Calabria
Contursi Terme	Salerno	Campania	San Lorenzo Nuovo	Viterbo	Lazio
Costigliole Salluzzo	Cuneo	Piemonte	San Mauro Castelverde	Palermo	Sicilia
Courmayeur	Valle d'Aosta	Valle d'Aosta	San Pellegrino Terme	Bergamo	Lombardia

Table 1. List of the 211 municipalities that were selected by the Town Attractiveness call, along with their province and region.

Municipality	Province	Region	Municipality	Province	Region
Crucoli	Crotone	Calabria	San Pietro di Cadore	Belluno	Veneto
Deliceto	Foggia	Puglia	San Romano in Garfagnana	Lucca	Toscana
Fabbriche di Vergemoli	Lucca	Toscana	Santa Croce Del Sannio	Foggia	Puglia
Falciano del Massico	Caserta	Campania	Santa Fiora	Agrigento	Sicilia
Fanano	Modena	Emilia-Romagna	Santa Lucia Del Mela	Benevento	Campania
Ferruzzano	Reggio Calabria	Calabria	Santa Severina	Grosseto	Toscana
Fobello	Vercelli	Piemonte	Santa Sofia	Messina	Sicilia
Foza	Vicenza	Veneto	Sant'Agata di Puglia	Crotone	Calabria
Gardone Riviera	Brescia	Lombardia	Sant'Angelo Muxaro	Forlì--Cesena	Emilia-Romagna
Gargnano	Brescia	Lombardia	Santo Stefano Quisquina	Agrigento	Sicilia
Genoni	Sud Sardegna	Sardegna	Sarmede	Treviso	Veneto
Ginestra	Potenza	Basilicata	Scilla	Reggio Calabria	Calabria
Gioia Sannitica	Caserta	Campania	Sellia	Catanzaro	Calabria
Goro	Ferrara	Emilia-Romagna	Selva di Cadore	Belluno	Veneto
Gradara	Pesaro e Urbino	Marche	Senales	Bolzano/Bozen	Trentino-Alto Adige
Gratteri	Palermo	Sicilia	Seneghe	Oristano	Sardegna
Grotte di Castro	Viterbo	Lazio	Sessa Cilento	Salerno	Campania
Guarene	Cuneo	Piemonte	Specchia	Lecce	Puglia
Isnello	Palermo	Sicilia	Sternatia	Lecce	Puglia
Isola del Gran Sasso d'Italia	Teramo	Abruzzo	Suvereto	Livorno	Toscana
Isola Dovarese	Cremona	Lombardia	Tarzo	Treviso	Veneto
Labro	Rieti	Lazio	Tolfa	Roma	Lazio
Laiuglia	Savona	Liguria	Tortorella	Salerno	Campania
Latera	Viterbo	Lazio	Tramonti	Salerno	Campania
Lenola	Latina	Lazio	Traona	Sondrio	Lombardia
Lizzano in Belvedere	Bologna	Emilia-Romagna	Triora	Padova	Veneto
Londa	Firenze	Toscana	Tribano	Imperia	Liguria
Lusiana	Vicenza	Veneto	Tripi	Messina	Sicilia
Malcesine	Verona	Veneto	Tusa	Messina	Sicilia
Marebbe	Bolzano/Bozen	Trentino-Alto Adige	Ucria, Raccuja e Floresta	Messina	Sicilia
Marradi	Firenze	Toscana	Uggiano la Chiesa	Lecce	Puglia
Massa Martana	Perugia	Umbria	Ustica	Palermo	Sicilia
Melpignano	Lecce	Puglia	Valentano	Viterbo	Lazio
Mesero	Milano	Lombardia	Vallecorsa	Frosinone	Lazio
Minervino di Lecce	Lecce	Puglia	Varzi	Pavia	Lombardia
Mompeo	Rieti	Lazio	Ventasso	Reggio nell'Emilia	Emilia-Romagna
Monasterace	Reggio Calabria	Calabria	Ventimiglia di Sicilia	Palermo	Sicilia
Moncalvo	Asti	Piemonte	Vernasca	Piacenza	Emilia-Romagna
Monchio delle Corti	Parma	Emilia-Romagna	Villachiara	Brescia	Lombardia
Monte Castello di Vibio	Perugia	Umbria	Villafrati	Palermo	Sicilia
Montecatini Val di Cecina	Pisa	Toscana	Villanova Monteone	Sassari	Sardegna
			Villaurbana	Oristano	Sardegna

Table 1 (continued)

5. Results

5.1 Extraction of town interventions

The first data extracted by ChatGPT was the list of interventions of the municipalities (or aggregations of) that were selected by the Town Attractiveness call described in the previous chapter. Table 2 shows such a list for the first four municipalities, in alphabetical order: Alano di Piave, Alcara li Fusi (aggregated with San Marco d'Alunzio), Amandola (aggregated with Montedinove and Rotella) and Ameno. Appendix 1 illustrates the same list, extended to 10 towns instead of 4. We didn't need to extract the list for all the 211 municipalities to examine all the possible issues, as after 10 towns it was clear that the same few issues repeated over and over.

Last two columns of table 2 show the type of errors in the data extraction and their reason. We identified three types of errors: wrong or missing budgets, missing interventions or swapped rows. The first type of error is related to the OCR conversion of the numbers that made up the budget of the intervention: sometimes, they may be wrongly recognized as different numbers or they can even be lost in the conversion, resulting in a missing budget. The second type of error is due to the skipping of a full table row by the OCR software; it usually happens when a table extends in more than one page: in this case, if the intervention is not listed in the first page of the table, there is a high probability that the algorithm doesn't recognize that the table continues in the next page, so it stops extracting interventions. This is the most common type of extraction error observed.

Finally, the third error type consists in swapping the budget with that of another intervention on a different row of the table ("swapped row"); such an issue is obviously due to the incorrect pairing of the elements in the rows of the table. This error is more subtle as at a first glance the budgets and the interventions look well extracted, but they are not.

Municipality	Proposed intervention	Budget (€)	ChatGPT's Budget	Error type	Cause
Alano di Piave	Efficientamento energetico MUST – Museo Storico Territoriale	351800	351800	none	
	Recupero del patrimonio storico: adeguamento funzionale, strutturale ed impiantistico di Villa Cinespa	697228	697228	none	
	Recupero del patrimonio storico: riqualificazione dello spazio pubblico collegato alla "Casetta dell'acqua"	50280	50280	none	
	Riqualificazione degli spazi pubblici aperti miglioramento dell'arredo urbano	18300	18300	none	
	Progetto di educazione, sensibilizzazione ed informazione giovanile	40138	40138	none	
	Realizzazione spettacoli ed eventi	30500	30500	none	
	Avvio alla gestione dell'infopoint	48830	48830	none	
	Realizzazione di percorsi turistici ciclo pedonali	100504	100060	different budget	OCR software
	Realizzazione delle infrastrutture di Villa Cinespa: infopoint, strutture ricettive e servizio di bikesharing	122060	67130	swapped row	new page
	Realizzazione della sezione primavera presso la scuola dell'infanzia di Alano di Piave	67130	73230	swapped row	new page
Progetti digitali per l'ammodernamento del sito internet e la gestione del turismo	73230	122060	swapped row	new page	
Alcara Li Fusi, San Marco d'Alunzio	Riqualificazione del percorso sentieristico intercomunale San Marco D'Alunzio	598559	598559	none	
	Realizzazione di servizi e infrastrutture culturali	316067	316067	none	
	Residenzialità diffusa	210652	210652	none	
	Borgo in musica	74722	74722	none	
	Notti al Castello	70000	70000	none	
	Digitalizzazione per la fruizione di beni culturali	85000	70000	swapped row	new page
	Digitalizzazione dei borghi	52000	168000	swapped row	new page
	Festa del Muzzuni	80000	85000	swapped row	new page
	Costruzione di una Cooperativa di comunità per la gestione dei servizi turistici e culturali	70000	52000	swapped row	new page
	Remote Experience e Art Residency	168000	80000	swapped row	new page
	Promozione iniziative imprenditoriali	90000		missing intervention	new page
	Head Quarter Village	65000		missing intervention	new page
	Intervento 14	50000		missing intervention	new page
	Intervento 16	35000		missing intervention	new page
Worldwide Culture	70000		missing intervention	new page	
Valorizzazione eccellenza produttiva	45000		missing intervention	new page	
AMANDOLA, Montedinove, Rotella	MUSEO PERGAMENE & COWORKING	250000	250000	none	
	TESORI DEL GUSTO: MARRONCINO	300000	300000	none	
	COWORKING E CENTRO INFO	220000	191409	wrong budget	OCR software
	SISTEMAZIONE ARCHIVIO E ISTITUZIONE CENTRO STUDI	130000	130000	none	
	COORDINAMENTO PARTNER E COMUNITÀ PATRIMONIALE (INTERVENTO TRASVERSALE)	95000	95000	none	
	FESTIVAL JAM SESSION (INTERVENTO TRASVERSALE)	100000	0	wrong budget	OCR software
	1 DIGITALIZZAZIONE E RETE MUSEI RAM (INTERVENTO TRASVERSALE)	100000	0	wrong budget	OCR software
	PERCORSO STRADA ROMANTICA DEI SIBILLINI (INTERVENTO TRASVERSALE)	100000	100000	none	
	RIQUALIFICAZIONE ARREDO URBANO E PANCHINA DELLA MELA ROSA DEI SIBILLINI	50000	100000	wrong budget	OCR software
	PISTA LUNGO LAGO SAN RUFFINO e N.2 PUNTI PANORAMICI STRADA ROMANTICA	150000	150000	none	
	REALIZZAZIONE N.4 PUNTI PANORAMICI STRADA ROMANTICA	130000	130000	none	
	AMPLIAMENTO CENTRO DI ACCOGLIENZA TURISTICA	370000	370000	none	
	CASA DELLE ASSOCIAZIONI E SCUOLA DI MUSICA	100000	100000	none	
CASA DELLE ASSOCIAZIONI	250000	250000	none		
SUPPORTO IMPRESE TURISTICHE E AGROALIMENTARI (INTERVENTO TRASVERSALE)	215000	215000	none		
Ameno	Quadrilatero della Cultura	750000	750000	none	
	Amena Libera Accademia	250000	250000	none	
	Residenze artistiche "Locus Amoenus"	160000	160000	none	
	Festival Multidisciplinare per lo spettacolo dal vivo	240000	240000	none	
	Educazione in natura e bosco civico	85000	85000	none	
	Accoglienza diffusa	77000	77000	none	
	Mobilità Green	80000	80000	none	
	Generazione Zero	42000	42000	none	
	Il Quadriborgo	75000	75000	none	
	Comunità di pratica	41300	41300	none	
	Coordinamento e amministrazione	195000	195000	none	

Table 2. List of interventions proposed by the first four municipalities selected by the Town Attractiveness call (in alphabetical order), and their associated budgets. The budget extracted by ChatGPT-4 is also shown, for a comparison with the real one. In case of issues with the data extraction of the intervention or of its budget, the type of error is specified, along with its reason.

Overall, ChatGPT was able to correctly extract all the data on the interventions and their budget only in case of one of the ten municipalities checked, the town of Ameno (province of Novara, Piedmont). In the case of Ameno, the table in Section 2 with the list of all interventions was split in two different pages, so it seems that the OCR software sometimes is able to recognize that the same table continues in a new page.

Only 61 interventions out of the 143 listed in Appendix 1 were correctly extracted (budget included), a mere 43% of the total. With such a low score, it is easy to conclude that at present ChatGPT-4 cannot be employed in a reliable way to select and extract data from complex documents.

Another reason that still prevent ChatGPT from analyzing very long pdf and/or a high number of pdf is that it can process only a finite number of input words, before starting “forgetting” the first ones it read; such a textual window is called “context window” and in case of ChatGPT-4, this limit is set in 128.000 tokens, equivalent to roughly 100.000 words (see Chapter 1) or 300 pages. 300 pages may seem a lot, but in practice are not. In our case, we need to upload the content of 211 pdf of 20-30 pages each, so the limit is reached soon, after the first 10-15 pdf.

The context window is very important because it is also employed to remember long conversations, that is to say, the full sequence of prompts that the users inserted along with all ChatGPT’s answers.

5.2 Validation of administrative information

It is also interesting to understand if ChatGPT is able to automatically generate the list of provinces and regions associated to the list of 211 municipalities shown in table 1. ChatGPT was trained on all the content available in the Wiki, so it should also know the exact province and region of each Italian municipality. We inserted the following simple prompt to generate the provinces and regions:

“Associate to each municipality in table 1 its province and its region, and insert these two new fields in the same table”

Table 1 was provided separately to ChatGPT as input data. In this case, the LLM is able to retrieve all the provinces and regions from its memory and it generates a table

identical to table 1.

We also compared the list of 211 municipalities provided by the MiC with the list of municipalities that ChatGPT extracted from the file name of the pdf (each pdf is called with the same name of the municipality it describes). In this case, the prompt is different:

“Associate to each municipality in table 1 its province and its region, taken from the provided spreadsheet. Insert these two new fields in the same table. Ignore distinction between uppercase and lowercase, between apostrophes and accents”

The spreadsheet cited in the prompt is a simple one, with just the names of the municipalities previously extracted by ChatGPT. In this case, the algorithm fail at pairing eight municipalities (table 3).

The reason is simple: accents. A few municipalities present an accent in their name, or an apostrophe at its end. In these cases, ChatGPT-4 strangely is not able to pair them to the same names in the spreadsheet, even if the only difference is the absence of accent or a different type of apostrophe (< ‘ > instead of < ' >) and even if the explicitly wrote in the prompt to not consider differences due to apostrophes and accents.

Municipality	Province	Region	Error reason
Arquà Petrarca	Padova	Veneto	Wrong accent
Casolino d'Erba	Como	Lombardia	Different < ' > symbol
Castelnuovo Bocca d'Adda	Lodi	Lombardia	Different < ' > symbol
Castrignano de' Greci	Ascoli Piceno	Marche	Different < ' > symbol
Isola del Gran Sasso d'Italia	Teramo	Abruzzo	Different < ' > symbol
Patù	Lecce	Puglia	Different < ' > symbol
Rocchetta Sant'Antonio	Foggia	Puglia	Different < ' > symbol
Sant'Agata di Puglia	Crotone	Calabria	Different < ' > symbol

Table 3. List of the eight municipalities that ChatGPT-4 failed at pairing.

This is interesting, as ChatGPT correctly recognized them when we asked to extract their region and provinces, even if also in that case accents were different from those stored in memory. Anyway, these results demonstrates that in same circumstances

ChatGPT is not as “smart” as we think, as any human would be able to understand that the town “Arqua Petrarca” is the same one of “Arquà Petrarca”. This is a practical example that proves that LLMs are not able to attribute a meaning to the words they so efficiently manipulate.

Finally, we noticed that strangely ChatGPT-4 is not able to extract geographical information from a list of municipalities we provided it (e.g: the list of latitude and longitude values of each town), even if this information in theory should be present in its training dataset. This is a very common operation, useful when working with GIS.

5.3 Error sources of ChatGPT-4

After many prompts to extract data from the 211 pdf, we identified several issues and classified them for type:

1. Intrinsic errors of ChatGPTs: these errors arises since LLMs are not able to give a meaning to words, as discussed in section 1.4. Thus, there is high probability that they will never be solved by any LLM, at least during this decade.
2. Technical limitations: these errors are due to simply technical constrains and will be probably overcome in the future versions of ChatGPT. Most of the errors identified belong to this category.
3. Choices of OpenAI: these are not real errors, but rather limitations that were imposed by OpenAI’s developers, for example in order to include more functionalities in the Premium version of ChatGPT.
4. Missing information: even if ChatGPT-4 was trained on a massive amount of data, it is not omniscient and sometimes ask the user to upload data the it needs to answer fully.

Table 4 shows each one of these error source, with some examples for each error type. In some cases, a solution was made available by OpenAI during the writing of this thesis (see column 3), or by using external software ad Open Interpreter or a VPN.

Error types	Examples	Solution
1. Intrinsic errors of ChatGPT that is not able to give a meaning to words	1. Some towns or hamlets may have the same name of other ones in different provinces or regions. At present, only a human is able to understand to which province they belong, reading the rest of the project description	Manually select each town
	2. The associations between some municipalities belonging to different lists or tables may be wrong because a different accent in the name of the municipality ("Not found" error)	Correct accents manually
2. Technical limitations of ChatGPT	1. ChatGPT4 is not able to load data and browse internet at the same time	Solved in update of November 2023
	2. ChatGPT-4 cannot load PDF of more than 30 MB	Bigger Context Window needed
	3. ChatGPT-4 cannot load .docx or .doc greater than 15 MB	Bigger Context Window needed
	4. ChatGPT-4 cannot load .txt of greater than 10 MB	Bigger Context Window needed
	5. Low processing speed and loading speed	Work outside office hours, install ChatGPT-4 with Open Interpreter
	6. ChatGPT4 often stop running after a while when time of execution is > 1 minute	OpenAI's servers need more computational power
	7. The number of prompts is limited to 50 each 3 hours	Use a VPN to connect to ChatGPT
	8. ChatGPT4 sometimes cannot load data at all, even if it correctly loaded the same file a few minutes before	No solution yet
	9. Lack of API do connect ChatGPT to all possible applications (e.g: email, PhotoShop, ...)	No solution yet
	10. Maximum number of files to upload is 10	Use GPT Store to bypass this limitation
	11. Sub-optimal OCR technology: ChatGPT often does not recognize that the same table can spread on more than one page of the document, and stop reading its data at the end of the first page of the table	Better OCR needed
3. Limitations imposed by choices of OpenAI	1. OpenAI decided to keep the feature of loading input files in the prompt only in ChatGPT-4	Use ChatGPT-4 instead
	3. You cannot tell ChatGPT to load data from your drive, and you have to upload it manually instead	Install Open Interpreter on your local machine
	4. ChatGPT is obliged to answer something after a while, even if it didn't finish the task assigned. However, doing so interrupts its task	No solution yet
4. Lack of enough information in the input data	1. We had to provide the list of italian municipalities	Solved in update of January 2024
	2. We had to provide the list of provinces and regions associated to each municipality	Solved in update of January 2024
	3. We had to provide the list of latitude and longitudes associated to each municipality, as the ones provided by ChatGPT-4 are not very accurate	No solution yet

Table 4. Main error sources of ChatGPT, divided by type, along with their possible solutions.

Two most needed advancements of ChatGPT-4 are its integration with third-party software and applications (e.g: email and social networks), that'd allow it to become a true virtual assistant, and the increase of its computational capacity, in order to serve millions of users simultaneously. Due to the widespread use of ChatGPT, in fact, it risks to collapse because of its own success, as the sheer number of users is actually slowing it down considerably.

At the time of writing, it is almost impossible to execute prompts that run for more than one minute. Prompts used to extract data are conceptually simple, but usually takes hours to run, as the number of documents to analyze is high (211 pdf, in our case). Thus, actually the only possible way to mine data from documents is to upload one of them at time in the graphical interface of ChatGPT-4, write the prompt to extract the desired information, save the data extracted and repeat the same process for the following document. This is a time-consuming task, as its length is often comparable to that of executing the same task manually, without the help of AI.

In case of analyzing hundreds of documents with ChatGPT-4, any prompt introduced still takes several hours of human work simply to run on all documents with the method

described above, not even considering the current maximum limit of 40 prompts per day of the Premium version. Relying on third-party software as Open Interpreter, that installs a version of ChatGPT-4 locally, is only a partial solution, as at present it does not come with the same advanced OCR software developed by OpenAI. This is not a technical issue but just a business one, as OpenAI is not interested in making its OCR technology freely available to the public.

Conclusions

The present work of thesis consists in two parts: the first two chapters are focused on the understanding of LLMs and of their various applications to the tourism sector. The novelty of this work consists in its second part, in which LLMs are employed to try to generate a touristic dataset, by extracting useful data from hundreds of documents published by the Ministry of Culture, in the framework of the recent Town Attractiveness Call of the PNRR.

Thanks to LLMs, the scientific community finally demonstrated on a practical basis what many neuroscientists thought for a long time: that biological memory resides in the strength of the synapses between neurons. Despite the fact that LLMs exhibit numerous emergent properties not anticipated by their developers, in our opinion AI won't replace humanity any soon, as LLMs are only statistical tools with an excellent memory, but unable to understand the very sentences they so skillfully generate.

Even if LLMs are not conscious algorithms, they are still tremendously useful and they are going to change forever the way we work and study: human workers won't be replaced by AI, but by other workers skilled in its use. Thus, University has to start teaching AI to students. After decades of slow advancements in the field of human-machines interactions, these sudden advancements allow every student and researcher to have a personalized AI assistant at his or her disposal, that significantly increases both efficiency and productivity. It is a revolution similar to the introduction of personal computers at the end of the XX century.

Between all possible applications of LLMs to tourism, we highlight three of the most important ones: improved language translations, that may open a whole new world of destinations to millions of tourists, personalization of the touristic experience to a level beyond what any travel agency could ever achieve, and enhancement of digital marketing strategies and planning.

The second goal of this work is about mining touristic data from the archives of PNRR tenders. At the time of writing (September 2024), popular LLMs as ChatGPT-4 are good in recognizing and extracting simple textual and numerical data stored in a variety of formats (e.g: .txt, .csv, .docx, .xlsx, .pptx, .html, .json, .pdf), but they fall behind in comprehending the layout of complex documents, particularly if they include many tables inside them.

Despite the best prompt that we could write, ChatGPT-4 was able to correctly extract only 43% of the town interventions described within the pdf of the project proposals. Such a low rate of success is not an intrinsic limitation of ChatGPT-4 itself, but of the external OCR technology that ChatGPT-4 employs to convert pdf in sequences of text: in case of complex pdf with tables and/or grid boxes, in fact, OCR software firstly converts each page in an image, and then it converts the images in text. Often, the conversion is sub-optimal and a certain percentage of the original text is lost or corrupted, making it impossible to analyze it further. This is particularly true in case of tables that span more than one page of the document, that are not recognized as part of the same table.

With the advancements of OCR software, which is now also based on AI, we foresee that these conversion issues will be partially solved in the near future, increasing the percentage of successful data retrieval from 43% to 95% or more, as it happened during the last decade in the case of non-generative AI tailored for object recognition. However, even with a success rate of 99%, AI won't be still useful to generate new datasets, as an error of 1% (e.g: 1 data every 100 entries) is too high to be acceptable. Instead, we should wait until the OCR algorithms reach the landmark of 99.99% of accuracy, as an error of 0.01% (e.g: 1 row every 10.000) is lower than average human accuracy. It might take a while then, before LLMs will be able to skillfully extract information from complex textual documents.

Until OpenAI won't solve or minimize this limitation, we don't suggest our Department to buy ChatGPT-4 licenses for research purposes. Future works of thesis will update these results in terms of measuring and reporting the improvements of LLMs in handling large and complex documents. They will also compare the skill of different LLMs in recognizing tables of text, as at the time of writing only ChatGPT-4 is able to upload and analyze pdf.

Appendix

The following table lists the interventions proposed in the framework of the Town Attractiveness tender of the PNRR. A total of 211 aggregations of municipalities were selected in all Italian regions, for an overall investment of 1 billion euro.

The results of both type of data extraction are compared in columns 3 and 4. The first column simply shows the name of the municipalities (or of their aggregations) described in the pdf. Only the first ten aggregations are shown in this table, as after analyzing a few aggregations all the errors of ChatGPT became evident.

Municipality	Proposed intervention	Budget (€)	ChatGPT's Budget	Error type	Cause
Alano di Piave	Efficientamento energetico MUST – Museo Storico Territoriale	351800	351800	none	
	Recupero del patrimonio storico: adeguamento funzionale, strutturale ed impiantistico di Villa Cinespa	697228	697228	none	
	Recupero del patrimonio storico: riqualificazione dello spazio pubblico collegato alla "Casetta dell'acqua"	50280	50280	none	
	Riqualificazione degli spazi pubblici aperti miglioramento dell'arredo urbano	18300	18300	none	
	Progetto di educazione, sensibilizzazione ed informazione giovanile	40138	40138	none	
	Realizzazione spettacoli ed eventi	30500	30500	none	
	Avvio alla gestione dell'infopoint	48830	48830	none	
	Realizzazione di percorsi turistici ciclo pedonali	100504	100060	different budget	OCR software
	Realizzazione delle infrastrutture di Villa Cinespa: infopoint, strutture ricettive e servizio di bikesharing	122060	67130	swapped row	new page
	Realizzazione della sezione primavera presso la scuola dell'infanzia di Alano di Piave	67130	73230	swapped row	new page
Progetti digitali per l'ammodernamento del sito internet e la gestione del turismo	73230	122060	swapped row	new page	
Alcara Li Fusi, San Marco d'Alunzio	Riqualificazione del percorso sentieristico intercomunale San Marco D'Alunzio	598559	598559	none	
	Realizzazione di servizi e infrastrutture culturali	316067	316067	none	
	Residenzialità diffusa	210652	210652	none	
	Borgo in musica	74722	74722	none	
	Notti al Castello	70000	70000	none	
	Digitalizzazione per la fruizione di beni culturali	85000	70000	swapped row	new page
	Digitalizzazione dei borghi	52000	168000	swapped row	new page
	Festa del Muzzuni	80000	85000	swapped row	new page
	Costruzione di una Cooperativa di comunità per la gestione dei servizi turistici e culturali	70000	52000	swapped row	new page
	Remote Experience e Art Residency	168000	80000	swapped row	new page
	Promozione iniziative imprenditoriali	90000		missing intervention	new page
	Head Quarter Village	65000		missing intervention	new page
	Intervento 14	50000		missing intervention	new page
	Intervento 16	35000		missing intervention	new page
	Worldwide Culture	70000		missing intervention	new page
Valorizzazione eccellenza produttiva	45000		missing intervention	new page	

Table 5. List of the proposed interventions and their budgets of the first 10 municipalities (or town aggregations) that participated to the 'Town attractiveness' call, in alphabetical order. The budget extracted with ChatGPT-4 is also shown and, in case of a discrepancy with the real value, the type of error is specified, along with the reason behind it.

Municipality	Proposed intervention	Budget (€)	ChatGPT's Budget	Error type	Cause
AMANDOLA, Montedivino, Rotella	MUSEO PERGAMENE & COWORKING	250000	250000	none	
	TESORI DEL GUSTO: MARRONCINO	300000	300000	none	
	COWORKING E CENTRO INFO	220000	191409	wrong budget	OCR software
	SISTEMAZIONE ARCHIVIO E ISTITUZIONE CENTRO STUDI	130000	130000	none	
	COORDINAMENTO PARTNER E COMUNITÀ PATRIMONIALE (INTERVENTO TRASVERSALE)	95000	95000	none	
	FESTIVAL JAM SESSION (INTERVENTO TRASVERSALE)	100000	0	wrong budget	OCR software
	1 DIGITALIZZAZIONE E RETE MUSEI RAM (INTERVENTO TRASVERSALE)	100000	0	wrong budget	OCR software
	PERCORSO STRADA ROMANTICA DEI SIBILLINI (INTERVENTO TRASVERSALE)	100000	100000	none	
	RIQUALIFICAZIONE ARREDO URBANO E PANCHINA DELLA MELA ROSA DEI SIBILLINI	50000	100000	wrong budget	OCR software
	PISTA LUNGO LAGO SAN RUFFINO e N.2 PUNTI PANORAMICI STRADA ROMANTICA	150000	150000	none	
	REALIZZAZIONE N.4 PUNTI PANORAMICI STRADA ROMANTICA	130000	130000	none	
	AMPLIAMENTO CENTRO DI ACCOGLIENZA TURISTICA	370000	370000	none	
	CASA DELLE ASSOCIAZIONI E SCUOLA DI MUSICA	100000	100000	none	
	CASA DELLE ASSOCIAZIONI	250000	250000	none	
SUPPORTO IMPRESE TURISTICHE E AGROALIMENTARI (INTERVENTO TRASVERSALE)	215000	215000	none		
Ameno	Quadrilatero della Cultura	750000	750000	none	
	Amena Libera Accademia	250000	250000	none	
	Residenze artistiche "Locus Amoenus"	160000	160000	none	
	Festival Multidisciplinare per lo spettacolo dal vivo	240000	240000	none	
	Educazione in natura e bosco civico	85000	85000	none	
	Accoglienza diffusa	77000	77000	none	
	Mobilità Green	80000	80000	none	
	Generazione Zero	42000	42000	none	
	Il Quadri borgo	75000	75000	none	
	Comunità di pratica	41300	41300	none	
Coordinamento e amministrazione	195000	195000	none		
Ampezzo, Preone, Socchieve	RIORDINO E DIGITALIZZAZIONE DELL'ARCHIVIO DI SOCCHIEVE	61000	61000	none	
	REALIZZARE INTERVISTE, FOTO E VIDEO CON IL COINVOLGIMENTO DEI CITTADINI	30000	Not Provided	missing budget	OCR software
	TEATRO NELLE SCUOLE COME STRUMENTO DI CONOSCENZA DEL PATRIMONIO MATERIALE E IMMATERIALE	90000	Not Provided	missing budget	OCR software
	DIGITALIZZAZIONE DELLA PINACOTECA DAVANZO (CON VIDEOSORVEGLIANZA)	70000		missing intervention	new page
	COMPLETAMENTO E ARREDO DI INFRASTRUTTURE TURISTICO RICETTIVE NEI COMUNI DI PREONE E SOCCHIEVE PER FAVORIRE L'ACCESSIBILITÀ	200000		missing intervention	new page
	INFOPOINT 4.0 AMPEZZO	100000		missing intervention	new page
	ESPERIENZA STORICO NATURALISTICA CON REALTÀ VIRTUALE PRESSO PALAZZO LUPIERI A PREONE	100000		missing intervention	new page
	ATTIVITÀ TEATRALI E MUSICALI A BENEFICIO DI RESIDENTI E VISITATORI CON IL COINVOLGIMENTO DELLE ASSOCIAZIONI LOCALI	120000		missing intervention	new page
	INFRASTRUTTURE E SERVIZI PER I CITTADINI (TELEMEDICINA, SERVIZI 0-3 ANNI, ETC)	100000		missing intervention	new page
	PRODURRE CONTENUTI MULTIMEDIALI SU PERCORSI, RICETTE, SEGRETI E PATRIMONIO CULTURALE MATERIALE E IMMATERIALE	40000		missing intervention	new page
	PIANO DI COMUNICAZIONE ONLINE E OFFLINE	99460		missing intervention	new page
	LA CULTURA DEL CIBO TRA TRADIZIONE E INNOVAZIONE	24400		missing intervention	new page
	EVENTI CULTURALI PER LA RIGENERAZIONE TERRITORIALE CON RESIDENZE ARTISTICHE	80000		missing intervention	new page
	MUSEO DIFFUSO DI ARTE	320000		missing intervention	new page
	SENTIERI STORICI DI COLLEGAMENTO TRA I BORGHI	45140		missing intervention	new page
	SELEZIONE ESPERTI PER ATTUAZIONE, MONITORAGGIO DEL PROGETTO (2 FIGURE X 36 MESI)	210000		missing intervention	new page
	SELEZIONE PERSONALE INTERNO PER L'ATTUAZIONE E LA RENDICONTAZIONE PER 36 MESI	50000		missing intervention	new page
	ALLESTIMENTO COWORKING E INCUBATORE D'IMPRESA A SOCCHIEVE	100000		missing intervention	new page
ACCOMPAGNAMENTO E FORMAZIONE PER LA CREAZIONE D'IMPRESA E FUNDRAISING PER IL TERZO SETTORE, PRESSO LA STRUTTURA DI SOCCHIEVE	120000		missing intervention	new page	

Table 5 (Continues)

Municipality	Proposed intervention	Budget (€)	ChatGPT's Budget	Error type	Cause
Antrodoto	Teatro S. Agostino: potenziamento dotazioni	49440	49440	none	
	Terme Spazio Polivalente - Danza e Circo Contemporaneo	353080	292000	wrong budget	OCR
	Palazzo Blasetti – Sala della Musica	535313	439981	wrong budget	OCR
	Polo Museale Antrodoto – Dotazioni e Spazi	131372	107772	wrong budget	OCR
	Residenza per Artisti	67830		missing intervention	
	Stagione Performing Arts	100830		missing intervention	
	Stagione della Musica	56830		missing intervention	
	Corso di formazione: musica, teatro, circo contemporaneo, danza	36600		missing intervention	
	Corsi di formazione: Tecnico del Suono e della Luce e Imprenditoria Culturale	18300		missing intervention	
	Ostello il Castagno – Potenziamento Ricettività	154220		missing intervention	
	Percorsi Montani e Urban Trekking - escursioni Eco-naturali e Artistico-Culturali	38674		missing intervention	
Comunicazione diffusa	57510		missing intervention		
Ardesio	Manutenzione straordinaria per Abbattimento BBAA di edificio comunale destinato a Museo E., Biblioteca e Uff. Comunali	57000	57000	none	
	Manutenzione straordinaria edificio Comunale da destinare a Biblioteca	376000	376000	none	
	Manutenzione straordinaria edificio Comunale per riqualificazione degli Uff. Comunali e degli spazi Museali	267000	267000	none	
	Interventi di efficientamento energetico (impianto di illuminazione pubblica) del Centro Storico del Borgo	109000	109000	none	
	Realizzazione Festival Cinematografico "Sacrae Scenae"	50000	Not Provided	missing budget	OCR software
	Realizzazione nuova area sosta camper in Via Frua	133000	133000	none	
	Manutenzione straordinaria ponte della Pista ciclopedonale	250000	250000	none	
	Riqualificazione area verde per Parco Giochi Inclusivo, in Via Locatelli	328000	328000	none	
	Costituzione Cooperativa di Comunità	10000	20000	wrong budget	row swap
Attività di Informazione, Promozione e Comunicazione	20000	10000	wrong budget	row swap	
Armungia	I nuovi luoghi della cultura: lo Spazio della Memoria della Comunità e lo Spazio della civiltà nuragica	620000	620000	none	
	Il Museo etnografico e il Museo Lussu come musei della comunità, innovativi e pienamente accessibili	75000	75000	none	
	Un nuovo polo culturale in Viale Gramsci: il Centro di Documentazione Storica, lo Spazio Formazione e la biblioteca comunale	205000	205000	none	
	I servizi culturali come strumento per la valorizzazione del patrimonio del paese	160000	160000	none	
	Immaginare il passato come risorsa per il futuro: la storia locale come patrimonio della comunità	90000	90000	none	
	Armungia come luogo di incontro, studio e progettazione culturale: i campus universitari	150000	150000	none	
	Armungia come luogo di formazione, arte e animazione: il progetto Cantieri Culturali	65000	45000	wrong budget	OCR software
	Dal paese al suo territorio: l'itinerario di collegamento tra i luoghi della cultura e la valorizzazione del Sentiero Italia CAI	50000	49180	none	
	Dare valore all'ospitalità: il sistema ricettivo di accoglienza Armungia hospitality	45000	65573	wrong budget	OCR software
	Per una cittadinanza temporanea: gli studenti dell'Università e i ricercatori come nuovi residenti	60000	65000	wrong budget	OCR software
Un piano di comunicazione per la promozione dell'offerta culturale	80000	50000	wrong budget	OCR software	

Table 5 (Continues)

Municipality	Proposed intervention	Budget (€)	ChatGPT's Budget	Error type	Cause
Arquà Petrarca	Realizzazione di Auditorium naturale entro il brolo di Casa del Petrarca	250000	250000	none	
	Creazione del percorso petrarchesco entro Arquà Petrarca	10000	10000	none	
	Studio del patrimonio artistico e archeologico di Arquà Petrarca, Catalogazione in SIGEC e database della Regione del Veneto	50000	50000	none	
	Restauro Oratorio SS. Trinità	200000		missing intervention	
	Manutenzione straordinaria Foresteria Callegari	260000		missing intervention	
	Manutenzione Fonte del Petrarca	100000		missing intervention	
	Manutenzione arredo urbano	200000		missing intervention	
	Censimento dei saperi immateriali di Arquà Petrarca	15000		missing intervention	
	Creazione di percorsi fisici di fruizione delle storie e dei saperi del borgo	10000		missing intervention	
	Ampliamento del portale web per la biblioteca perduta del borgo	5000		missing intervention	
	Ologrammi, videomapping, ricostruzioni virtuali & digital events	70000		missing intervention	
	Iniziative dedicate alla scrittura, poesia e musica con le scuole di Arquà Petrarca	25000		missing intervention	
	Creazione dell'hub delle associazioni entro la Foresteria Callegari	25000		missing intervention	
	Ampliamento della biblioteca	10000		missing intervention	
	Creazione di una piattaforma digitale di gestione della cultura ad Arquà Petrarca	30000		missing intervention	
	Accompagnamento del progetto	150000		missing intervention	
	Ampliamento del sito internet dedicato ad Arquà Petrarca con l'inserimento di nuove lingue (francese, spagnolo)	2500		missing intervention	
	Servizio con veicolo navetta elettrico tra aree di sosta e borgo (spese per l'avvio)	5000		missing intervention	
	Consolidamento Infopoint con brochures e pubblicazioni	5000		missing intervention	
	Abbattimento delle barriere architettoniche fisiche e immateriali	10000		missing intervention	
	Potenziamento delle infrastrutture per la mobilità lenta e sostenibile	27500		missing intervention	
	Creazione della programmazione di eventi dedicati al borgo e al Petrarca	5000		missing intervention	
	Eventi: primo test sul coordinamento pubblico-privato per anniversario del Petrarca e relativa analisi	100000		missing intervention	
	Spese di comunicazione a mezzo stampa, web, social e con influencers e testimonials	10000		missing intervention	
Creazione di mostre sui saperi materiali e immateriali del borgo	10000		missing intervention		
Tavoli di lavoro con Comune di Padova, OgD, Parco Colli, Regione, Università e privati coinvolti	0		missing intervention		
Collegamenti con parchi europei e spese di missione per collegamenti con Parchi Letterari e con siti del Petrarca in Italia e Francia	5000		missing intervention		
Arsiero	RISTRUTTURAZIONE EDIFICIO "PALAZZO DI CARTA"	1200000	1200000	none	
	ARREDAMENTO BIBLIOTECA, SPAZI STUDIO, UFFICIO TURISTICO E PERTINENZE "PALAZZO DI CARTA"	100000	100000	none	
	ALLESTIMENTO MUSEALE "PALAZZO DI CARTA" E CREAZIONE PERCORSO TEMATICO DI ARCHEOLOGIA INDUSTRIALE	75000	75000	none	
	CREAZIONE DI UN UFFICIO DI PROMOZIONE TURISTICA DEL TERRITORIO E DELLE TRADIZIONI	20000	20000	none	
	DESTINATION MANAGEMENT ORGANIZATION	10000	30000	wrong budget	row swap
	FESTIVAL ANNUALE DELL'ACQUA E DELLA CARTA	30000		missing intervention	OCR software
	CREAZIONE ARCHIVIO STORICO COMUNALE E DIGITALIZZAZIONE DEI DOCUMENTI	75000	75000	none	
	ACQUISTO DI MEZZI PER UNA MOBILITA' TURISTICA SOSTENIBILE	20000	20000	none	
	AVVIO ACCADEMIA MUSICALE PER GIOVANI	30000	30000	none	
	COMUNICAZIONE E DIVULGAZIONE SPAZIO E INIZIATIVE "PALAZZO DI CARTA"	30000	30000	none	

Table 5 (Continues)

The second column of Table 5 illustrates the name of the interventions. ChatGPT-4 sometimes did not retrieve one or more interventions, but when it did, it always extracted the correct names present inside the pdf, without any misspelling. The third column lists the budget associated to each intervention, manually taken from the table in Section 2 of the project proposal; in this case, there may be some extraction errors in the numbers of the budgets; thus, column 4 was introduced in order to compare the budgets extracted by humans with those extracted by ChatGPT-4.

A few recurrent error types were identified and specified in column 5. They are described in section 5.1, “Extraction of town interventions”, along with the sources of these errors, highlighted in column 6.

Acronyms

AI: Artificial Intelligence is a broad concept, usually related to algorithms able to simulate capacities typically attributed to human or animal minds, such as thought or object recognition

AGI: Artificial General Intelligence is a generalization of the concept of AI. Such an hypothetical AI would also be able to give a *meaning* to the input it manipulates (text, images or audio), thus possessing a certain degree of consciousness

ANN: it stands for Artificial Neural Networks, a common type of AI based on algorithms that try to simulate the behavior of biological neurons

ChatBot: a mean of communication between humans and machines, usually in the form of a graphical interface with a box, in which humans can insert their prompts, to ask questions to the machines

ChatGPT: a particular type of ChatBot, where the machine is powered by a LLM called GPT. It was developed by OpenAI and launched on November 30, 2022

DNN: Deep Neural Networks are a particular type of ANN, designed to have at least six layers of artificial neurons. ChatGPT is based on a DNN with about 100 layers.

Generative AI: any type of AI whose main aim is to generate data, typically text or images.

GPT: Generative Pre-trained Transformer. It is the software architecture of the LLM developed by OpenAI since June 2018. Pre-trained means that in order to work, the model was given an input in the form of a massive amount of text and images taken

from the web, that it used to calibrate its parameters. Transformer is the name of the particular technique employed to process data, in order to be the most efficient and fast possible

GPU: Graphic Processor Unit. A specialized type of processors that enable fast graphical rendering of images and video, and it also greatly increase computational speed of LLMs with the Transformer architecture, as GPT

LLM: Large Language Model. A kind of generative AI made up by a DNN and designed to output sequences of text

MiC: acronym of the Italian Ministry of Culture

OCR: Optical Character recognition. A software designed to recognized handwritter or typed text in an image, and convert it into a sequence of words. Recently, AI has significantly enhanced OCT technology.

OpenAI: the developers of the LLMs known as GPT, later paired to a ChatBot and presented al ChatGPT

PNRR: National Recovery and Resilience Plan is Italy's strategic response to the global COVID-19 emergency, addressing the challenges that have hindered the country's economic, social, and environmental development over the year. It forms a crucial part of the Next Generation EU, European Union's recovery mechanism

Bibliography

- Baggio, R., & Yildirim, P. (2021). Regulating artificial intelligence: A literature review of current legal and ethical implications. *Journal of Responsible Technology*, 9, 100027.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).
- Bonera, G. (2023). Sustainable tourism and regeneration of Italy's hamlets: comparative analysis of two PNRR projects in Lombardy
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism Management*, 29(4), 609-623.
- Buhalis, D., & Sinarta, Y. (2019). Real-time co-creation and nowness service: Lessons from tourism and hospitality. *Journal of Tourism Futures*, 5(1), 7-23.
- Camilleri, M. A. (2018). The use of data-driven technologies for marketing: A critical research agenda. *Journal of Hospitality and Tourism Technology*, 9(3), 289-301.
- Cheng, F. F., & Jin, X. (2012). Adoption of mobile technology in tourism: A stage-based perspective. *Tourism Management*, 33(5), 1197-1207.
- Cheong, R. (1995). The virtual threat to travel and tourism. *Tourism Management*, 16(6), 417-422.
- Choi, H. S. C., & Morrison, A. M. (2008). Increasing the attractiveness of destination web sites: A qualitative evaluation of photographs. *Journal of Travel Research*, 46(2), 172-185.

- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303-314.
- Dale, H. H. (1935). Chemical transmission of the effects of nerve impulses. *British medical journal*, 1(3874), 835.
- Dann, G. M. S. (1993). Limitations in the use of 'nationality' and 'country of residence' variables. *Tourism Marketing and Management Handbook*, 459-467.
- Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q. V., ... & Ng, A. Y. (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 1223-1231.
- Dinan, C., & Sargeant, A. (2000). Social marketing and sustainable tourism—is there a match? *International Journal of Tourism Research*, 2(1), 1-14.
- Doe, J., et al. (2019). Dynamics of the Tourism Industry: A Data-Driven Approach. *International Journal of Hospitality Management*, 45, 77-89.
- Dolnicar, S. (2008). Market segmentation in tourism. *Tourism Management*, 29(1), 1-15.
- Dowling, R. (2013). *Sustainable Tourism and Cultural Heritage*. Routledge.
- Drachman, D. A. (2005). Do we have brain to spare? *Neurology*, 64(12), 2004-2005.
- Edgell, D. L., Allen, L. R., & Smith, K. A. (2008). *Tourism policy and planning: Yesterday, today, and tomorrow*. Routledge.
- Enz, C. A., & Thompson, G. M. (2013). The best of times, the worst of times: Differences in hotel performance following 9/11. *Cornell Hospitality Quarterly*, 54(4), 325-341.
- Evans, R., & Patel, M. (2020). The Rise of Large Language Models in AI. *AI & Society*, 35(3), 567-578.
- Gitelson, R. J., & Crompton, J. L. (1983). The planning horizons and sources of information used by pleasure vacationers. *Journal of Travel Research*, 21(3), 2-7.
- Goodwin, H. (2011). Tourism, responsibility, and sustainability: Reflecting on the challenge of measuring progress. *Journal of Sustainable Tourism*, 19(2), 169-184.
- Gössling, S., et al. (2020). Artificial intelligence and sustainable tourism: The road ahead. *Journal of Sustainable Tourism*, 28(8), 1123-1137.
- Green, P. (2022). Big Data and AI: The Future of Tourism. *Journal of Digital Tourism*, 4(2), 200-215.
- Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of Tourism Research*, 38(3), 757-779.

- Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of Tourism Research*, 38(3), 757-779.
- Gretzel, U. (2018). Influencer marketing in travel and tourism. *Tourism Management Perspectives*, 25, 151-155.
- Gretzel, U., & Fesenmaier, D. R. (2010). Harnessing the wisdom of crowds in travel. *Tourism and Hospitality Research*, 10(1), 73-77.
- Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In *Information and Communication Technologies in Tourism 2008* (pp. 35-46). Springer, Vienna.
- Gretzel, U., et al. (2021). The future of AI in tourism: A roadmap for research. *Information Technology & Tourism*, 23(1), 1-29.
- Gretzel, U., Fesenmaier, D. R., Formica, S., & O'Leary, J. T. (2015). Searching for the future: Challenges faced by destination marketing organizations. *Journal of Travel Research*, 45(2), 116-126.
- Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179-188.
- Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179-188.
- Guttentag, D. A. (2010). Virtual reality: Applications and implications for tourism. *Tourism Management*, 31(5), 637-651.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
- Hall, C. M., Gossling, S., & Scott, D. (Eds.). (2015). *The Routledge handbook of tourism and sustainability*. Routledge.
- Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Prentice Hall PTR.
- Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. New York: Wiley.
- Heilweil, Rebecca (December 7, 2022). "AI is finally good at stuff. Now what?". Vox. Archived from the original on January 16, 2023. Retrieved December 30, 2022.
- Hertz, G. (1887). Ueber den Einfluss, welchen die Stromstärke auf die Erregung der Nerven durch den galvanischen Strom ausübt. *Archiv für die gesammte Physiologie des Menschen und der Thiere*, 40(1), 460-482.
- Hu, Krystal (February 2, 2023). "ChatGPT sets record for fastest-growing user base – analyst note". Reuters. Archived from the original on February 3, 2023. Retrieved June 3, 2023.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168-177.

- Huang, Y. C., Backman, S. J., Backman, K. F., & Chang, L. L. (2016). Exploring the implications of virtual reality technology in tourism marketing: An integrated research framework. *International Journal of Tourism Research*, 18(2), 116-128.
- Hudson, S., & Thal, K. (2013). The impact of social media on the consumer decision process: Implications for tourism marketing. *Journal of Travel & Tourism Marketing*, 30(1-2), 156-160.
- Ivanov, S., & Webster, C. (2017). Adoption of robots, artificial intelligence and service automation by travel, tourism and hospitality companies – A cost-benefit analysis. *International Journal of Contemporary Hospitality Management*, 29(9), 2481-2500.
- Jannai, D., Meron, A., Lenz, B., Levine, Y. & Shoham, Y. Preprint at <https://arxiv.org/abs/2305.20010> (2023).
- Johnson, M., Smith, K., Jones, L., & Williams, K. (2023). Pushing the Boundaries: Anticipated Features of ChatGPT-5. *AI and Robotics Symposium Proceedings*, 45-58.
- Jones, D. (2023). Empowering Researchers with AI. *Advances in Computer Science*, 47(1), 34-50.
- Jung, T., Lee, H., Chung, N., & Tom Dieck, M. C. (2018). Cross-cultural differences in adopting mobile augmented reality at cultural heritage tourism sites. *International Journal of Contemporary Hospitality Management*, 30(3), 1621-1645.
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (2013). *Principles of Neural Science*. McGraw-Hill Education.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59-68.
- Kavoura, A. (2019). Social media platforms as marketing tools in tourism: A review. In *Marketing Places and Spaces*, 135-148.
- Khan, M., & Ali, R. (2019). Virtual Assistants and Programming Languages. *Journal of Computer Science*, 15(11), 1594-1605.
- Kim, M. J., Lee, C. K., & Jung, T. (2020). Exploring consumer behavior in virtual reality tourism using an extended stimulus-organism-response model. *Journal of Travel Research*, 59(1), 69-89.
- Lakshmanan, Lak (December 16, 2022). "Why large language models like ChatGPT are bullshit artists". becominghuman.ai. Archived from the original on December 17, 2022. Retrieved January 15, 2023.
- Law, R., Leung, R., & Buhalis, D. (2009). Information technology applications in hospitality and tourism: A review of publications from 2005 to 2007. *Journal of Travel & Tourism Marketing*, 26(5-6), 599-623.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing*, 30(1-2), 3-22.

- Li, F., & Zhang, Y. (2021). Paradigm Shifts in AI: An Overview. *Journal of Artificial Intelligence Research*, 69, 45-76.
- Liu, Z., & Wall, G. (2006). Sustainability in tourism: A review of the literature. *Journal of Sustainable Tourism*, 14(6), 629-648.
- Lu, W. (2014). Google Translate Service. *Journal of Chinese Language and Computing*, 24(1), 52-56.
- Luo, J., & Zhong, H. (2020). AI in tourism: A systematic literature review. *Journal of Travel Research*, 59(3), 509-523.
- Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Mao, Z., Ding, Y., & Zhang, L. (2015). Understanding travelers' e-word-of-mouth intentions on social media in tourism: An exploratory study from a mixed methods perspective. *Journal of Travel & Tourism Marketing*, 32(7), 851-868.
- Marcus, G. (2018). Deep Learning: A Critical Appraisal. arXiv preprint arXiv:1801.00631.
- Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management*, 34, 99-107.
- McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 3-44). MIT Press.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115-133.
- Miller, R., & Davis, S. (2018). The Role of Social Media in Tourism Marketing. *Travel Research Journal*, 29(3), 112-128.
- Moro, S., Rita, P., & Vala, B. (2019). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 96, 334-347.
- Morrison, A. M. (2010). *Marketing and managing tourism destinations*. Routledge.
- Munar, A. M., & Jacobsen, J. K. S. (2014). Motivations for sharing tourism experiences through social media. *Tourism Management*, 43, 46-54.
- Nature. (2022). What 'Next Generation Italia' means for research. *Nature*. Retrieved December 6, 2023, from <https://www.nature.com/articles/d43978-021-00005-w>
- OpenAI (November 30, 2022). "ChatGPT: Optimizing Language Models for Dialogue". Archived from the original on November 30, 2022. Retrieved December 5, 2022

- Pan, B., et al. (2020). Understanding tourists through location-based social media data: A literature review. *Journal of Travel Research*, 59(6), 1027-1043.
- Pang, A. (2014). Managing the media in times of crisis: A case analysis of the media strategies employed by Resorts World Sentosa and Singapore Tourism Board in the 2010 casino crises. *Journal of Hospitality and Tourism Management*, 21, 100-109.
- Pearce, P. L. (1993). Fundamentals of tourist motivation. In D. G. Pearce & R. W. Butler (Eds.), *Tourism Research: Critiques and Challenges* (pp. 113-134). Routledge.
- Peterson, G. (2021). GPT and Its Applications in Data Mining. *Data Science Review*, 6(2), 89-101.
- Pike, S. (2002). Destination image analysis—a review of 142 papers from 1973 to 2000. *Tourism Management*, 23(5), 541-549.
- Proserpio, D., & Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, 36(5), 645-665.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). Improving Language Understanding by Generative Pretraining. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
- Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 873-880).
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
- Ramón y Cajal, S. (1899). *Textura del sistema nervioso del hombre y los vertebrados* (Vol. 1). Imprenta y Librería de Nicolás Moya.
- Reich, Aaron (December 27, 2022). "ChatGPT: What is the new free AI chatbot? – explainer". The Jerusalem Post. Archived from the original on January 18, 2023. Retrieved December 30, 2022.
- Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12), 46-54
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2020). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(5), 1-37.
- Richards, G., & Munsters, W. (Eds.). (2010). *Cultural tourism research methods*. CABI.
- Rider, Elizabeth (April 6, 2023). "How ChatGPT Will Dramatically Change the Influencer Space". Entrepreneur. Archived from the original on April 13, 2023. Retrieved April 25, 2023.

- Roivainen, E., 2023 I gave ChatGPT an IQ test. Here's what I discovered. <https://www.scientificamerican.com/article/i-gave-chatgpt-an-iq-test-heres-what-i-discovered/>
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- Scheyvens, R., & Biddulph, R. (2017). Responsible tourism: Conceptualisation, diversity and implications. *Routledge Handbook of Tourism and Sustainability*, 143-158.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Shepherd, G. M. (1998). The synaptic organization of the brain. *Oxford University Press*, USA.
- Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1), 141-159.
- Sigala, M. (2018). New technologies in tourism: From multi-disciplinary to anti-disciplinary advances and trajectories. *Tourism Management Perspectives*, 25, 151-155.
- Sigala, M., et al. (2021). AI in tourism and hospitality: A state-of-the-art review. *Journal of Travel Research*, 60(1), 3-20.
- Smith, J., & Johnson, A. (2020). Challenges in Tourism Research Data Management. *Journal of Tourism Studies*, 31(2), 45-59.
- Song, H., & Li, G. (2019). Predicting tourism demand with composite search index. *Journal of Travel Research*, 58(8), 1391-1405.
- Sparks, B. A., So, K. K. F., & Bradley, G. L. (2016). Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, 53, 74-85.
- Sui, Xiubao & Wu, Qiu hao & Liu, Jia & Chen, Qian & Gu, Guohua. (2020). A Review of Optical Neural Networks. *IEEE Access*. PP. 1-1. 10.1109/ACCESS.2020.2987333.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Taylor, L., & White, G. (2017). Overcoming Data Analysis Challenges in Tourism Research. *Tourism Analysis*, 22(1), 131-143.
- Tung, Liam (January 26, 2023). "ChatGPT can write code. Now researchers say it's good at fixing bugs, too". ZDNET. Archived from the original on February 3, 2023. Retrieved June 22, 2023.
- Turing, A. M. *Mind* LIX, 433–460 (1950).
- Tussyadiah, I. (2020). A review of research into automation in tourism: Launching the Annals of Tourism Research Curated Collection on artificial intelligence and robotics in tourism. *Annals of Tourism Research*, 81, 102883.

- Tussyadiah, I. P., & Pesonen, J. (2016). Impacts of peer-to-peer accommodation use on travel patterns. *Journal of Travel Research*, 55(8), 1022-1040.
- Tussyadiah, I. P., & Zach, F. J. (2017). Identifying salient attributes of airline choice: A comprehensive review and future research. *Journal of Travel & Tourism Marketing*, 34(4), 534-555.
- Tussyadiah, I. P., Jung, T. H., & tom Dieck, M. C. (2018). Embodiment of wearable augmented reality technology in tourism experiences. *Journal of Travel Research*, 57(5), 597-611.
- Tussyadiah, I. P., Wang, D., & Jia, C. H. (2017). Exploring the persuasive power of virtual reality imagery for destination marketing. *Journal of Travel Research*, 56(1), 128-139.
- Valls, J. F., Boger, C. A., & Freitag, R. C. (2000). Application of electronic commerce in the tourism industry: A case study of electronic travel market. *Journal of Travel Research*, 39(2), 164-171.
- Varanasi, Lakshmi (January 5, 2023). "ChatGPT creator OpenAI is in talks to sell shares in a tender offer that would double the startup's valuation to \$29 billion". Insider. Archived from the original on January 18, 2023. Retrieved January 18, 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (pp. 5998-6008).
- Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., ... & Cao, Y. (2019). Artificial intelligence in marketing: Opportunities and challenges. *Journal of the Academy of Marketing Science*, 48(4), 617-641.
- Wang, X., & Chung, N. (2022). Textual and Graphical Data Analysis in Tourism.
- Wei, W. (2019). Investigating the impact of using virtual reality tours in marketing on tourists' satisfaction and loyalty. *Journal of Vacation Marketing*, 25(3), 337-351.
- Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.
- Werthner, H., & Klein, S. (1999). Information technology and tourism: A challenging relationship. *Springer Science & Business Media*.
- Xiang, Z., et al. (2021). Artificial intelligence in tourism and hospitality: A review of current literature and future research directions. *Journal of Hospitality and Tourism Technology*, 12(3), 491-508.
- Xiang, Z., Magnini, V. P., & Fesenmaier, D. R. (2015). Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet. *Journal of Retailing and Consumer Services*, 22, 244-249.
- Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 44, 120-130.

Xie, K. L., & Johns, N. (1999). Service failure and recovery: The impact of relationship factors on customer satisfaction. *Journal of Service Research*, 2(4), 350-370.

Yao, Y., Grinberg, E., & Wallace, B. C. (2020). Challenges and opportunities in automated machine learning. *Nature Machine Intelligence*, 2(1), 29-31.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2), 634-639.

Zheng, X., et al. (2019). Predicting tourist arrivals at points of interest using big data. *Journal of Travel Research*, 58(4), 537-551.

Zomerdijk, L. G., & Voss, C. A. (2010). Service design for experience-centric services. *Journal of Service Research*, 13(1), 67-82.