

Article

Not peer-reviewed version

Crafting a Museum Guide Using GPT4

[Georgios Trichopoulos](#)*, [Markos Konstantakis](#), [George Caridakis](#), [Akrivi Katifori](#), Myrto Koukouli

Posted Date: 22 June 2023

doi: 10.20944/preprints202306.1618.v1

Keywords: Museum; Cultural Heritage; Digital Storytelling; GPT4; Artificial Intelligence



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Crafting a Museum Guide Using GPT4

Georgios Trichopoulos ^{1,*}, Markos Konstantakis ^{2,†}, George Caridakis ^{3,‡},
Akrivi Katifori ^{4,‡} and Myrto Koukouli ^{5,‡}

¹ University of the Aegean; gtricho@aegean.gr

² University of the Aegean; mkonstadakis@aegean.gr

³ University of the Aegean; gcari@aegean.gr

⁴ University of Athens; vivi@di.uoa.gr

⁵ University of Athens; mkoukouli@di.uoa.gr

* Correspondence: gtricho@aegean.gr

† Current address: University of the Aegean, Department of Cultural Technology and Communication, University Hill, 81132, Mytilene, Greece

‡ Current address: National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, Panepistimiopolis, 15771, Ilissia, Athens, Greece

Abstract: This paper introduces a groundbreaking approach to enriching the museum experience using GPT4, a state-of-the-art language model by OpenAI. By developing a museum guide powered by GPT4, we aim to address the challenges visitors face in navigating vast collections of artifacts and interpreting their significance. Leveraging the model's natural language understanding and generation capabilities, our guide offers personalized, informative, and engaging experiences. However, caution must be exercised as the generated information may lack scientific integrity and accuracy. To mitigate this, we propose incorporating human oversight and validation mechanisms. The subsequent sections present our own case study, detailing the design, architecture, and experimental evaluation of the museum guide system, highlighting its practical implementation and insights into the benefits and limitations of employing GPT4 in the cultural heritage context.

Keywords: museum; cultural heritage; digital storytelling; GPT4; artificial intelligence

1. Introduction

The rapid advancements in artificial intelligence (AI) have opened exciting possibilities across various domains, and the field of cultural heritage is no exception. Computational methods are used in digital storytelling which is seamed to any kind of museum visit [1]. Museums are developing chatbots to assist their visitors and to provide an enhanced visiting experience [2]. In this paper, we present a groundbreaking approach to enhancing the museum experience using GPT4, a state-of-the-art language model. Museums have long served as repositories of human knowledge and cultural heritage, providing visitors with a unique opportunity to explore history, art, and science [3,4]. However, navigating through vast collections of artifacts and interpreting their significance can be a challenging task, especially for visitors with limited background knowledge [5,6]. Our research aims to bridge this gap by developing a museum guide powered by GPT-4, which leverages the model's natural language understanding and generation capabilities to offer personalized, informative, and engaging experiences for museum visitors. GPT-4, the latest iteration of OpenAI's renowned language model, represents a significant leap forward in AI-driven natural language processing. Building upon the successes of its predecessors, GPT-4 boasts enhanced contextual understanding, improved coherence, and a broader knowledge base [7,8]. Leveraging these advancements, our proposed museum guide seeks to revolutionize the way visitors interact with museum exhibits, providing them with a dynamic and immersive journey through history and culture. By tapping into GPT-4's vast corpus of information and its ability to generate human-like responses, our guide aims to offer tailored recommendations, detailed explanations, and captivating narratives, ultimately enriching the visitor's understanding and appreciation of the displayed artifacts. Through this novel integration of AI technology, we envision

a future where museum visits become more accessible, engaging, and intellectually rewarding for visitors of all backgrounds and interests. One of the remarkable qualities of GPT-4, the language model at the heart of our museum guide project, is its remarkable adaptability and versatility. Despite its vast knowledge base and the immense amount of information it has absorbed, GPT-4 can be tailored to serve specific purposes, such as acting as a personalized museum guide. This ability stems from the model's sophisticated architecture and its capacity for fine-tuning. By training GPT-4 on a curated dataset of museum-related content, we can effectively shape its responses and guide its behavior to align with the specific requirements of a museum guide application. This tailoring process ensures that the language model not only possesses a deep understanding of general cultural knowledge but also acquires a contextual understanding of the specific museum exhibits, allowing it to provide accurate and relevant information to visitors. This adaptability empowers GPT-4 to seamlessly integrate into the museum environment, making it a versatile and invaluable tool for enhancing the visitor experience and promoting a deeper appreciation of our shared heritage. While the use of GPT-4 as a museum guide holds great promise, it is crucial to approach the information generated by the language model with a certain degree of caution. As impressive as its capabilities may be, GPT-4 is still a machine learning model trained on vast amounts of data, including text from the internet. This reliance on pre-existing data introduces potential challenges related to the accuracy and scientific integrity of the information provided [9,10]. Firstly, GPT-4's responses are generated based on patterns and associations observed in its training data, rather than on true understanding or critical evaluation [11,12]. While efforts are made to curate the dataset used for fine-tuning the model, it is impossible to guarantee that all the information it absorbs is entirely accurate or up to date. Consequently, there is a risk that the language model might inadvertently propagate misconceptions, inaccuracies, or outdated knowledge to museum visitors. Secondly, GPT-4 lacks the ability to verify the authenticity or reliability of the information it generates [13]. Unlike human experts who can critically analyze and cross-reference multiple sources, the language model does not possess the capacity for independent fact-checking. Therefore, there is a need for human oversight to ensure the veracity of the information provided by the museum guide. To address these concerns, it is essential to incorporate robust validation mechanisms and maybe human supervision in the deployment of the museum guide. Expert curators and domain specialists should work in collaboration with GPT-4, reviewing and verifying the information generated by the model to ensure its accuracy and scientific integrity. By combining the strengths of AI technology with human expertise, we can strive to deliver a museum guide that provides reliable and trustworthy information while acknowledging the inherent limitations of machine-generated knowledge. In the subsequent sections of this paper, we delve into the development and evaluation of our own museum guide system, built upon the foundation of GPT-4. Through a case study and experimentation, we aim to showcase the practical implementation of our proposed approach, highlighting its strengths and identifying potential challenges. We present the design and architecture of the museum guide, detailing the methodologies employed to fine-tune GPT-4 for the specific context of cultural heritage. Moreover, we provide insights into the data collection process, curation techniques, and the integration of human expertise to ensure the reliability and accuracy of the information presented to museum visitors. By sharing our experiences and findings, we contribute to the ongoing discourse on the use of AI technology in the cultural sector and offer valuable insights into the potential benefits and limitations of employing GPT-4 as a museum guide. The following text goes as follows: Section 2 describes the state-of-the-art works in the field, on Section 3 we describe our own system called MAGICAL and we conclude with results on Section 4. Section 5 is the References.

2. Related Work

Over the past two years, the rapid evolution of Generative Pre-trained Transformers (GPT) has had a profound impact across various sectors, revolutionizing the way we interact with technology. One area that has been significantly influenced by GPT's advancements is the field of cultural heritage research. With its ability to understand and generate human-like text, GPT has opened new avenues

for exploring and preserving our rich cultural past. By analyzing vast amounts of historical data, manuscripts, artworks, and artifacts, GPT will become an invaluable tool for researchers, enabling them to gain deeper insights into different aspects of cultural heritage. This technology has not only accelerated the process of digitizing and cataloging artifacts but has also enhanced our understanding of ancient civilizations, languages, and traditions. By bridging the gap between artificial intelligence and cultural heritage, GPT has become an indispensable asset in unlocking the secrets of our collective history. At the time of writing, the latest version of GPT is 4 and looks clearly improved during the tests compared to previous versions. It produces complete texts, without semantic and syntactical errors, without repetitions and ambiguities. It is a new technology, and it will take some time for researchers of all fields to discover the full range of possibilities that this tool has given. Bubeck et al. [14] try to discover this new potential, while Chang et al. [15] investigate which books are already known to GPT4. Siu et al. [16] explore the capabilities given to professional language translators and Chen et al. [17] are also researching on the language handling abilities and speech recognition of GPT4. In every culture there are stereotypes about genders, races, groups of people etc. and Cheng et al. [18] try to measure these stereotypes inside Large Language Models (LLMs), like GPT4. On the same path, Jiang et al. [19] investigate the ability of GPT to express personality traits and gender differences. Additionally, there are studies that discuss the potential implications of GPT in intellectual property and plagiarism [20], as well as the limitations and challenges of GPT models and their learning mechanisms [21]. Other studies focus on the use of advanced techniques in art conservation [22], on-site interpretative and presentative planning for cultural heritage sites [23], and the development of a thesaurus in an educational web platform on optical and laser-based investigation methods for cultural heritage analysis and diagnosis [24]. Therefore, while there is limited research on the use of GPT in cultural heritage applications, there are related studies that may provide insights into the potential applications and challenges of GPT in this field.

3. MAGICAL: Museum AI Guide for Augmenting Cultural Heritage with Intelligent Language model

MAGICAL's goal is to create a digital tour guide for any museum or cultural space that will be able to dialogue with the visitor, be able to suggest routes based on user response, and be able to create digital narratives, with real or fictional characters, with the aim of greater engagement and emotional involvement of the visitor. The digital tour guide can change the spoken language at any time - it always answers in the language asked or alternatively we can ask it to answer in the language we want. Communication with MAGICAL is done by natural speech. There is no need for the visitor to type text or read a screen. In this way we avoid the phenomenon of smartphone zombies [25–28], where site visitors end up glued to their mobile screen, losing contact with the real space. For this communication to be technically possible we need some wearable smart IoT device, and the closest existing technology is smart glasses. The glasses have an earpiece and a microphone, they allow the visitor to have their hands free and move without distracting the visual attention from the exhibits and the space around. In addition, they will allow some augmented content to be displayed if this is required later, in a future implementation. The digital tour guide can be adapted to any museum in a very easy way. It can change its speech style depending on the user's age and can be trained using real cultural data extracted from previous research projects or collected by museum curators. This makes it obvious that the tour guide could work very easily in any other cultural or non-cultural space. It is a system adaptable to any condition and this is a very great strength of GPT4.

3.1. MAGICAL System Architecture

The architecture of the system is simple and is illustrated in Figure 1. Conceptually, it can be divided into four modules. In the first module is the user, the user's interface with the system through the smart glasses and the user experience resulting from this communication. Communication with the tour guide starts from the user who will speak and possibly greet the guide or introduce himself.

The speech is converted to text in the second module, which is referred to in the diagram as the input module. Speech-to-text conversion, referred to by the initials ASR (Automatic Speech Recognition) or STT (Speech-To-Text) and the reverse process called TTS (Text-To-Speech), are already an advanced research area [29–36], and there are many functional, valuable tools available. OpenAI's Whisper [37], is a STT tool that works perfectly with GPT4 and can understand speech exceptionally, at any speed, as long as it's in English. Google has been providing its text-to-speech and speech-to-text API to developers through the Google Cloud Platform for several years. These services are provided free of charge for small-scale use or come with a fee depending on the volume of data traffic. It is possible to configure the voices that are heard for example the tone (male or female voice), the playback speed and the punctuation. The Google application works with almost all known languages and has evolved over the years into a reliable and functional tool. If a developer wants to create an STT service on its own, without having to pay a provider, there are Python libraries that work quite well. One of them is Coqui STT ¹, "an open-source deep-learning toolkit for training and deploying speech-to-text models", as described into the website. Once the speaker's voice is converted to text, it is sent to the GPT4 engine which responds and returns its response back in text form. GPT4 is a stateful engine, meaning it records the entire history of a conversation and reprocesses it before responding. Thus, it appears to have memory and be able to process and respond based on previous dialogue sentences. When these lines were written, the engine had not been given by OpenAI publicly but only to developers, in some order of priority. The reason is the excessively increased demand for the system which made the response from the servers very slow. The GPT runs on a distributed system whose capabilities have often been overwhelmed by the enormous number of queries it receives at any time from users worldwide. Depending on the workload of the system, usually the response from GPT4 has a small delay. After getting the text from the language model, we are interested in converting it back to audio. The responses from the model come in JSON text format and should feed the third module in the series, the one in Figure 1 called the Output Module. A Text-To-Speech engine should take care of converting text, in any language, to audio. English is one language that has been studied more than any other, but the real challenge is converting text from other, less-used languages to sound as close as possible to natural language. It is a very common phenomenon, the voice that sounds like a robot, without a particular timbre and without expressiveness. Google Cloud services work quite well for most of the spoken languages, and they are free for small-scale use. Another option which is completely free for developers is a Python library that works quite well. It is called pyttsx ², The original version works up to Python 2, version 3 works up to Python version 3, and now there is also the latest version of the library, the pyttsx4. It works without internet connection and delays and supports multiple TTS engines like espeak and Sapi5. Another service that works well for over 80 languages is SpeechGen.io ³ which provides its own API for developers. Of course, there are more STT and TTS tools available to developers. The sound of the answer should reach the user's smart glasses and a new cycle will start when the user starts to speak again.

¹ <https://stt.readthedocs.io/en/latest/#>, accessed on 15/06/2023

² <https://pypi.org/project/pyttsx4/>, accessed on 15/06/2023

³ <https://speechgen.io/>, accessed on 15/06/2023

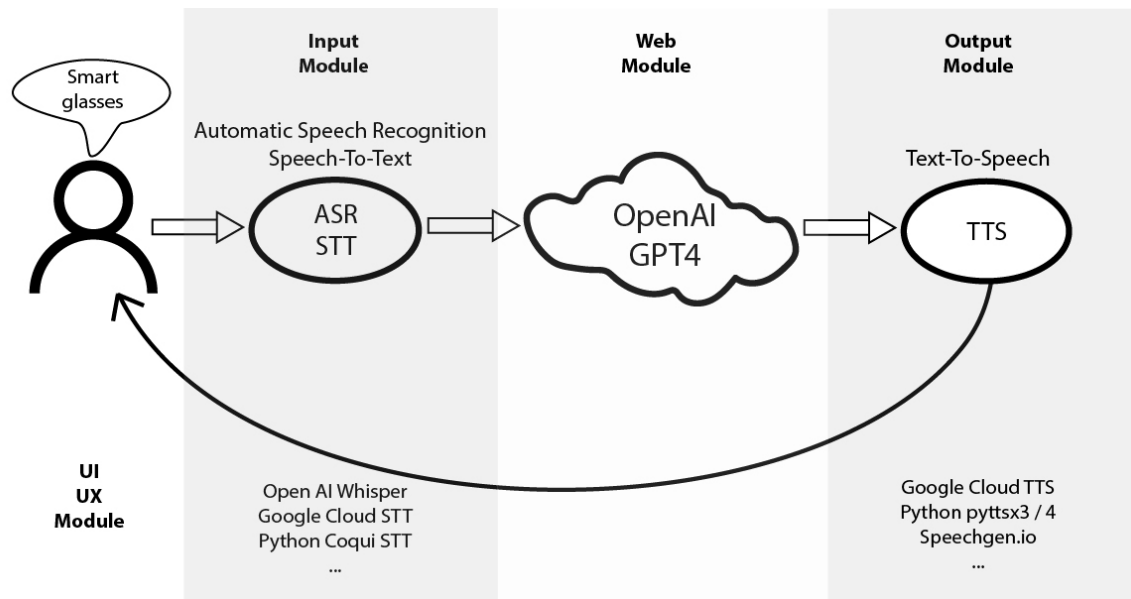


Figure 1. The MAGICAL System Architecture

3.2. Case Study – Chat with Ebutius and Calle

To test the operation of MAGICAL we turned to the researchers of the Narralive research team, who created the Narralive Storyboard Editor (NSE) and the Narralive Mobile Player (NMP) applications [38]. The NSE is a software tool to create narratives for cultural heritage while the NMP is a mobile app for the end-users, in which they can experience these narratives. The Narralive team has tested these tools in the Hunterian Museum in Glasgow, in a permanent exhibition about the Antonine Wall. The Antonine Wall was the northernmost frontier of the Roman Empire, and it was abandoned from the late AD150s, when the Romans retreated further south. In the exhibition, visitors can find many artifacts discovered along the Wall. The Narralive team created digital narratives in the form of audio, for the visitors of the exhibition. The purpose of the narratives was “to increase visitors’ engagement and connection with the objects on display, and more broadly with related themes, historic periods, heritage, museums and the past”. These narratives were created as part of the project called Emotive [39], and the exhibition was called “Ebutius’ dilemma” [40]. For the needs of the narratives, two fictional characters were used: Ebutius and Calle. Ebutius is a Roman army officer, a centurion, sent to serve at the frontier. Calle is the woman that Ebutius loved. She belongs to the tribe of the Caledonians. The Caledonians were the natives of Scotland before the arrival of the Romans. They were at war with the Romans, protecting their land. Thus, the stories raise issues of love, family, work, loss of loved ones, engaging the listener emotionally. They also stereotype relations between Romans and Caledonians and military life. So, the challenge was to use these stories and train the language model to know Ebutius and Calle, to learn their stories and be able to use them, without revealing that these people are fictional personas. When the first experiments began, there was GPT3. It was a language model that lacked the capacity for dialogue (stateless model). It was trained using generative pre-training, with a vast amount of data found on the internet. It could already answer questions like “What is Antonine’s wall?”, or “What is the northernmost tip of the Roman empire?”, but of course it knew nothing about Ebutius and Calle. Thus, a way had to be found for the information of the stories to enter the model. OpenAI gives this possibility to developers, at some cost. The original data of the stories, as received by the Narralive team had the JSON format. They consisted of 155 questions and 55 answers. Each answer can answer more than one question, but each question has a unique answer. For each answer, there was a set of questions in the following form:

“answer”: “The Romans conquered lands that the Caledonians considered their own, so many of them are justifiably angry at the Romans. Raids and skirmishes from the Caledonian

tribes were, in fact, a regular event. Nevertheless, some Caledonians co-existed rather peacefully with the Romans and traded with them frequently. For example, local style pottery was found in various forts, which indicates that there were local crafts people and merchants interacting with the army on the Wall. Also, soldiers of various ranks often married local women, although these marriages were not recognised by the Roman State until after Antoninus's rule (AD 138-161). After Antoninus's reforms, any children the soldiers might have had with these women were encouraged to join the Roman army and hence gain citizenship for themselves.",

"questions": [

"Could a Roman soldier marry a local Caledonian woman?",

"Why did the Caledonians attack the Romans?",

"What were the relationships between Romans and Caledonians?",

"How could Calle fall in love with Ebutius, the conqueror of her people's lands?",

"How was the relationship between Romans and the locals?",

"Did the natives complain about or disturb the building of the rampart?",

"Where and how did you meet your wife Calle?",

"Did the Caledonians interact with the Roman Army in the Wall?",

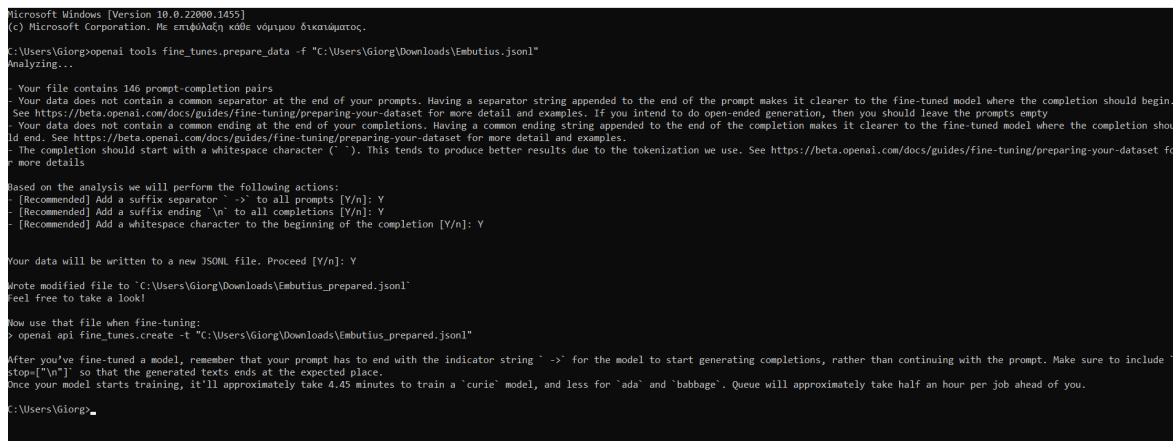
"Are the Caledonians angry at the Romans?"]

The first edit of the questions and answers found field names on each line that were not needed in this research. So, editing (renaming, removing fields, deleting duplicate records) was required. To fine tune a model in GPT3, it is necessary to prepare the data in a specific way. For this purpose, a tool called "CLI data preparation tool" (Figure 2) is provided, which can accept as input data in various formats (json, jsonl, xlsx, csv, tsv) (related instructions on the OpenAI's web page ⁴) and outputs a result.jsonl file. In our case, the tool did not work satisfactorily and the errors it displayed were difficult to debug. Thus, it was preferred to manually transfer the data to a new jsonl type file, with the formatting required by the OpenAI engine. The file ended up with 146 questions and answers. The text, after being manually edited, was successfully passed through the CLI preparation tool, which added spaces and special characters as required for GPT3 to function properly. In the final format, each line looked like this:

"prompt":"Were the Caledonians a Celtic tribe?" ->,"**completion":**" Yes. The Caledonians were a Celtic tribe that inhabited the areas of modern-day Scotland during the Roman era. They were builders and farmers and defeated and were defeated by the Romans on several occasions. Nearly all the information available about the Caledonians is based on predominantly Roman sources, which may suggest bias. During the Iron Age, Scotland did not have a nucleated settlement pattern. Instead, Caledonians lived in homesteads dispersed across the landscape, each occupied by an extended family and their dependents.\n"

That means that every line had a "prompt" and a pairing "completion", two concepts that are fundamental for the fine tuning of GPT3.

⁴ <https://beta.openai.com/docs/guides/fine-tuning>, accessed on 15/06/2023



```

Microsoft Windows [Version 10.0.22000.1425]
(c) Microsoft Corporation. Με επιφύλαξη νόμιμου δικαιώματος.

C:\Users\Giorg>openai tools fine_tunes.prepare_data -f "C:\Users\Giorg\Downloads\Ebutius.json"
Analyzing...

- Your file contains 146 prompt-completion pairs
- Your data does not contain a common separator at the end of your prompts. Having a separator string appended to the end of the prompt makes it clearer to the fine-tuned model where the completion should begin. See https://beta.openai.com/docs/guides/fine-tuning/preparing-your-dataset for more detail and examples. If you intend to do open-ended generation, then you should leave the prompts empty
- Your data does not contain a common ending at the end of your completions. Having a common ending string appended to the end of the completion makes it clearer to the fine-tuned model where the completion should end. See https://beta.openai.com/docs/guides/fine-tuning/preparing-your-dataset for more detail and examples.
- The completion should start with a whitespace character (' '). This tends to produce better results due to the tokenization we use. See https://beta.openai.com/docs/guides/fine-tuning/preparing-your-dataset for more details

Based on the analysis we will perform the following actions:
- [Recommended] Add a suffix separator ' ->' to all prompts [Y/n]: Y
- [Recommended] Add a suffix ending '\n' to all completions [Y/n]: Y
- [Recommended] Add a whitespace character to the beginning of the completion [Y/n]: Y

Your data will be written to a new JSONL file. Proceed [Y/n]: Y

Wrote modified file to 'C:\Users\Giorg\Downloads\Ebutius_prepared.json'
Feel free to take a look!

Now use that file when fine-tuning:
> openai api fine_tunes.create -t "C:\Users\Giorg\Downloads\Ebutius_prepared.json"

After you've fine-tuned a model, remember that your prompt has to end with the indicator string ' ->' for the model to start generating completions, rather than continuing with the prompt. Make sure to include 'stop:["\n"]' so that the generated texts ends at the expected place.
Once your model starts training, it'll approximately take 4.45 minutes to train a 'curie' model, and less for 'ada' and 'babbage'. Queue will approximately take half an hour per job ahead of you.

C:\Users\Giorg>

```

Figure 2. Using the CLI data preparation tool

The first tests were disappointing (Figure 3). The model does not always correctly answer the question “who is Ebutius”. It presents serious problems in its use such as truncated answers in 50-70 characters, controversial answers, and answers in the form of a new question. Sometimes the answer is the question itself. For the latter, the reason is that the model was trained on a set of prompts and completions, and since the prompts were in the form of questions, the model responded similarly. Also, a big problem was that the model training data sample was too small. OpenAI advises training with at least 1000 sets of prompts and completions. Of course, something like this would have cost a lot and on the other hand, it is difficult to find reliable and accurate cultural data of these sizes. In addition, using GPT through the Windows command prompt by calling the appropriate commands was not practical and easy. We needed a new user interface and more data for training the model. For the continuation of the tests, a simple GUI was built that allows text to be entered into a box and returns the answers of the language model. Other cultural data from internet sources were also searched and a dataset from the Kaggle.com site was used. The data set was 1155 rows, one for each listed UNESCO cultural heritage site for the year 2021. The amount of data was too large (and therefore expensive) to use all of it in training the model, and on the other hand, there was a very high probability, the language model was already fed with this data and any action on our part would be pointless. Anyway, two of the fifteen columns of data for each monument were used. The first two columns of data entitled Name and Description of the monument were kept for all 1155 rows of the table and were fed to the GPT3. Already from the first test of the new model there seemed to be a qualitative difference in the results obtained. They were clearly more targeted, and the model seemed to be able to describe any monument we were referring to. Of course, wrong answers appeared again. The GUI also helped the test flow a lot but the inability of the model to engage in conversation with the user made things difficult. We had to find a way to make the model function stateful. It should be able to remember all the previous dialogue. So, we started sending the GPT not only the last sentence but the whole dialogue from the beginning. That’s where the other weakness of the model appeared: The limitation on the number of tokens it can accept as input. The default limit is 2048 tokens and can go up to a maximum of 4096 tokens. The way OpenAI counts tokens is special - not fixed for every word and sentence. In general, a token corresponds to one syllable, except for very common and used words which may correspond entirely to a single token. So, the tests improved, and some form of dialogue actually started, but it could only last 2-3 sentences and then the model would go into a state where it would generate uncontrollable and meaningless texts. The solution to the dialog problem came directly from the creators of the model: they gave the developers the GPT3.5 version. It is a stateful model, the engine behind the very popular ChatGPT. It could again accept data through a very different training process (more like instructions to the model) and was additionally capable of complete dialogues, very close to natural human language. At its first version worked very well with English but had great difficulty with any other language. Beyond the capacity for dialogue the new model was clearly

upgraded in text production: errors were few or simply absent. One obvious observation was that the text might be repeating some meaning. Also, the generated text could be interrupted again without explanation, or conversely, be too long as if the model was babbling.

```

Microsoft Windows [Version 10.0.22000.1455]
(c) Microsoft Corporation. Με επιφύλαξη κάθε νόμιμου δικαιώματος.

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Who is Embutius?"
Who is Embutius? Don't our rulers have to be male in order to rule?

No

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "What is the Antonine wall?"
What is the Antonine wall? > The Antonine wall is the farthest north the Roman army managed to reach

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Is Embutius a roman soldier?"
Is Embutius a roman soldier? - Yes, of course. We think that he was a legionnaire of the

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Is Embutius a roman soldier? Please don't cut your phrase"
Is Embutius a roman soldier? Please don't cut your phrase lengths to 80 characters. Lengthy texts give a better idea of the everyday lives

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Is Embutius a roman soldier?"
Is Embutius a roman soldier?

No, he is not. We hope that all citizens of the Roman

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Do you know Embutius?"
Do you know Embutius? A No. We suspect that he was a close relative of Marcus Aurelius

C:\Users\Giorg>openai api completions.create -m davinci:ft-personal-2023-01-12-18-51-36 -p "Why did the Caledonians attacked the Romans?"
Why did the Caledonians attacked the Romans? That's a very good question. It might have been a straight forward raid for

C:\Users\Giorg>

```

Figure 3. The first tests of trained GPT3

The model improvements don't just stop at the ability to generate text. It is much less expensive to use (about 1/10 the cost of GPT3) and much simpler to guide. It uses some parameters with which the programmer gives it a role - what it is and how it should behave - and it can also accept instructions through sentences. Trying to use all 146 sentences from Narralive's data resulted in 4096 tokens limit exceeding error which also remains in this version. Even with the use of less data, the model responded impressively to dialogues in which it assumed the role of the Hunterian Museum curator (Figure 4). We proceeded with the tests by putting the system in place of Ebutius himself. It responded excellently but could not - despite the clear instruction it had received - hide the fact that Ebutius is a fictional character merely used to make the visit more pleasant: GPT cannot lie, describe someone in a bad way, use prejudice and discrimination of any kind. While our digital tour guide was still being tested, OpenAI gave us access to the API of GPT4. Within a very short time, we have seen the rapid development and improvement of the language model at all levels. With almost no changes to our code, no modification to the model training process, we were immediately ready for new tests and the results were even more impressive. Our tour guide can now speak languages other than English. Tested with Greek, French and Italian and was responsive and error free. It responds every time in the same language in which the user writes or can - if requested - translate into any language. The answers it gives to each question are more complete, they don't stop without reason, they contain more complete meanings. Also, the 4096 tokens limitation is now gone. The model accepted without a problem all 146 sets of prompts and completions and can, with the same ease as it was in GPT3.5, assume the role we will assign to it, in the style we will describe. The only drawback that the GPT4 shows compared to the previous models, is the lower speed of its response. To solve the problem OpenAI expanded its collaboration with Microsoft and in January 2023 announced the building of new multiple supercomputing systems to support the increased workloads⁵.

⁵ <https://openai.com/blog/openai-and-microsoft-extend-partnership>, accessed on 15/06/2023

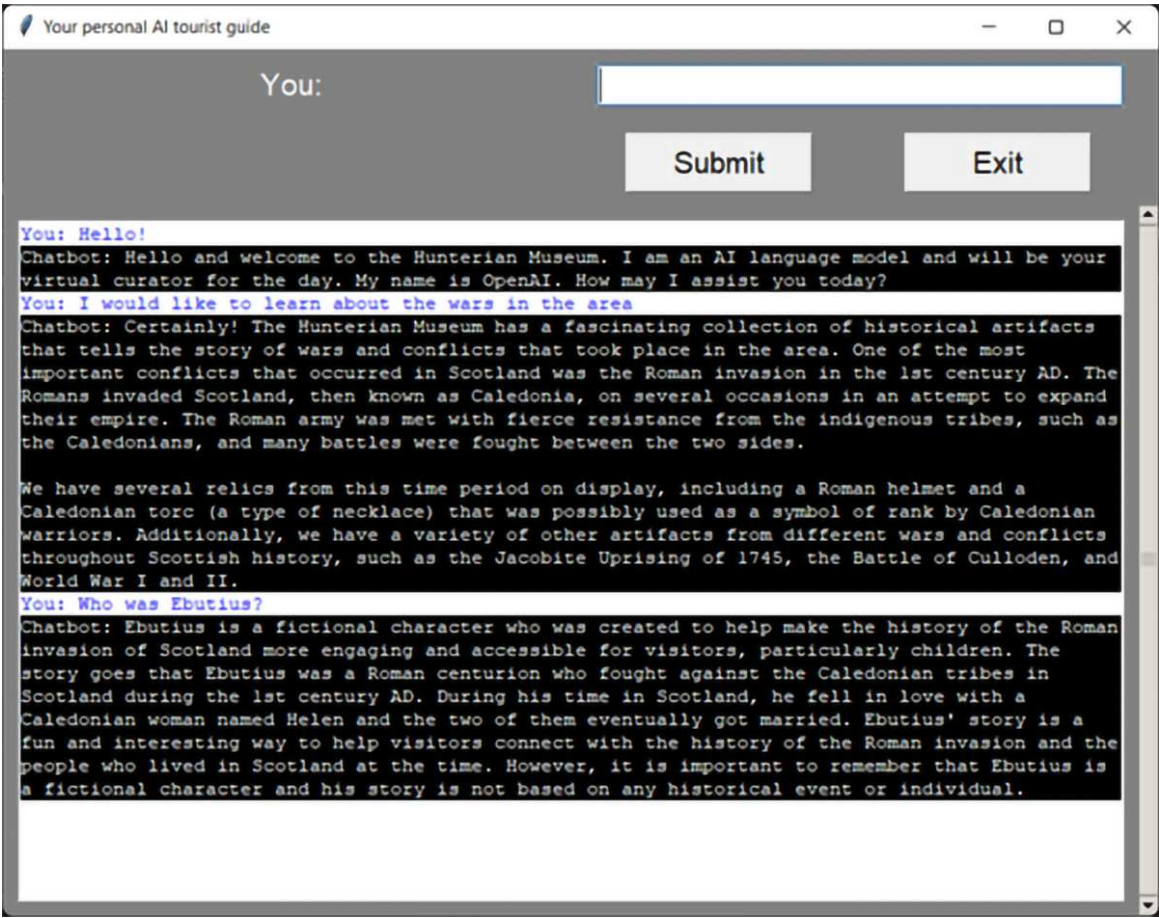


Figure 4. Testing of GPT3.5 using a GUI.

Table 1 summarizes what was written above regarding the characteristics of the GPT in its various versions, and its behavior during the testing of MAGICAL. We observe a continuous improvement in the functioning of the GPT language model: Its results are more and more convincing in terms of their plausibility, the use of the model has become easier and more affordable, and the potential given to the user is increasingly greater. On the other hand, this improvement was accompanied by a massive increase in the number of users that experiment with the model. The system was advertised by word of mouth with incredible speed and caused waves of excitement worldwide. This overload of the OpenAI’s servers led to a significant descend of their response.

Table 1. Comparison between versions of OpenAI GPT during tests.

No	Characteristic	GPT3	GPT3.5	GPT4
1	Can be fine-tuned	Yes	Yes	Yes
2	Bias in text	No	No	No
3	Ease in guidance	Low	High	High
4	Can change style of the text	No	Yes	Yes
5	Truncated answers	Yes	No	No
6	Extra-long answers (babbling effect)	No	Yes	No
7	Can use other languages than English	No	Partially	Yes
8	Repeated meanings	Yes	Yes	No
9	Controversial answers	Yes	Yes	No
10	Input tokens limitation	2048 (normal) - 4096 (max)	4096	None
11	Cost for training and use	High	Low	Low
12	Speed in responses	High	Medium	Low

4. Conclusion, and future works

From September 2022 to May 2023, the development of the GPT model was rapid. Day by day users worldwide have been given tools and capabilities that have never been given before. Artificial intelligence made leaps and bounds and entered the daily lives of millions of people, but they were not ready to take advantage of it and do not yet know the full range of possibilities that have been given to them. So, on the one hand there is an intense enthusiasm for all that has come and that we expect to be developed, but on the other hand we should probably be somewhat cautious. In the field of cultural heritage where we refer to museum exhibits, cultural sites and pieces of our history, there is not much room for errors and misinterpretations in the information that will be presented to the public. The language model can do an amazing job of talking to museum visitors, but can we ensure that it always feeds users with correct, scientifically valid, and evidenced knowledge? It's too early to know and it will take a lot of testing before we can put the finished product into use. We could though suggest some workable solutions. It would be perfectly safe, for example, to create an educational application - a tour guide for school students in which, after first the children being prepared for an exhibition by their teachers, and studying the space and exhibits they will visit, they evaluate the application, and they try to identify possible tour guide errors. This would make the experience of using the digital guide more playful while also keeping students engaged. Another safe idea would be to integrate MAGICAL into some existing, already tested application that creates cultural narratives. One such example could be the Narralive app. As the expected results from the existing application are known, the texts produced by the digital guide can be directly compared in terms of their validity. Again, the profit will be double as we will evaluate the results of MAGICAL while at the same time we will give a boost to the older app, which will be renewed and reused. To use the MAGICAL digital tour guide directly in a museum, it would certainly require supervision by an experienced curator for a period of use. The language model creates texts that look correct to the untrained eye but focus on detail is needed and testing before a finished product is released. In our next steps, we first intend to implement the input and output modules of Figure 1, which will be

connected to the existing application. It is necessary to communicate with the language model through natural speech. In addition, all functionality of the application should be connected to a mobile IoT device, which will leave the hands free. The Vuzix Blade smart glasses have been chosen as the test device because of their availability. In addition, we will be experimenting with the NAO robot from Softbank Robotics, which we have available. We will try to integrate the MAGICAL application into the robot and test it as a tour guide, thus continuing our research also in the field of tangible interfaces for producing digital narratives [41]. After all our testing, there will be a period of external user testing and evaluation of MAGICAL. The results will be published.

Author Contributions: Conceptualization, G.T.; methodology, G.T.; software, G.T.; validation, G.T., M.Kon.; investigation, G.T.; data curation, A.K., M.Kou., G.T.; writing—original draft preparation, G.T.; writing—review and editing, G.T., M.Kon., A.K., M.Kou; supervision, G.C.; project administration, G.T., G.C.; funding acquisition, M.Kon. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GPT	Generative Pre-trained Transformer
CH	Cultural Heritage
TTS	Text-to-speech
STT	Speech-to-text
ASR	Automatic Speech Recognition
AI	Artificial Intelligence
LLM	Large Language Model
API	Application Programming Interface
JSON	Javascript Object Notation
AD	Anno Domini
GUI	Graphical User Interface
UI	User Interface
UX	User Experience

References

1. Trichopoulos, G.; Alexandridis, G.; Caridakis, G. A Survey on Computational and Emergent Digital Storytelling. *Heritage* **2023**, *6*, 1227–1263. doi:10.3390/heritage6020068.
2. Varitimiadis, S.; Kotis, K.; Pittou, D.; Konstantakis, G. Graph-Based Conversational AI: Towards a Distributed and Collaborative Multi-Chatbot Approach for Museums. *Applied Sciences* **2021**, *11*. doi:10.3390/app11199160.
3. Lawan, S. Challenges and Prospect of Museum Institutions in the 21st Century in Northern Nigeria. *Journal of Social Sciences Advancement* **2022**, *3*, 45–52. doi:10.52223/JSSA22-030105-31.
4. Farahat, B.I.; Osman, K.A. Toward a new vision to design a museum in historical places. *HBRC Journal* **2018**, *14*, 66–78, [https://doi.org/10.1016/j.hbrj.2016.01.004]. doi:10.1016/j.hbrj.2016.01.004.
5. Carnall, M.; Ashby, J.; Ross, C. Natural history museums as provocateurs for dialogue and debate. *Museum Management and Curatorship* **2013**, *28*, 55–71, [https://doi.org/10.1080/09647775.2012.754630]. doi:10.1080/09647775.2012.754630.
6. Buchanan, S.A. Curation as Public Scholarship: Museum Archaeology in a Seventeenth-Century Shipwreck Exhibit. *Museum Worlds* **2016**, *4*, 155 – 166. doi:10.3167/armw.2016.040112.

7. Adesso, G. Towards The Ultimate Brain: Exploring Scientific Discovery with ChatGPT AI. **2023**. doi:10.22541/au.167052124.41804127/v2.
8. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. **2023**. doi:10.20944/preprints202303.0422.v1.
9. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **2023**, *3*, 121–154. doi:https://doi.org/10.1016/j.iotcps.2023.04.003.
10. Currie, G.M. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? *Seminars in Nuclear Medicine* **2023**. doi:https://doi.org/10.1053/j.semnuclmed.2023.04.008.
11. OpenAI. GPT-4 Technical Report. 2023, [arXiv:cs.CL/2303.08774].
12. Lehman, J.; Gordon, J.; Jain, S.; Ndousse, K.; Yeh, C.; Stanley, K.O. Evolution through Large Models. 2022, [arXiv:cs.NE/2206.08896].
13. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; Wu, Z.; Zhu, D.; Li, X.; Qiang, N.; Shen, D.; Liu, T.; Ge, B. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. 2023, [arXiv:cs.CL/2304.01852].
14. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M.T.; Zhang, Y. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023, [arXiv:cs.CL/2303.12712].
15. Chang, K.K.; Cramer, M.; Soni, S.; Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. 2023, [arXiv:cs.CL/2305.00118].
16. Siu, S.C. ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. *Available at SSRN 4448091* **2023**.
17. Chen, F.; Han, M.; Zhao, H.; Zhang, Q.; Shi, J.; Xu, S.; Xu, B. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. 2023, [arXiv:cs.CL/2305.04160].
18. Cheng, M.; Durmus, E.; Jurafsky, D. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. 2023, [arXiv:cs.CL/2305.18189].
19. Jiang, H.; Zhang, X.; Cao, X.; Kabbara, J. PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences. 2023, [arXiv:cs.CL/2305.02547].
20. Dehouche, N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics* **2021**, *21*, 17–23.
21. Lee, M. A Mathematical Interpretation of Autoregressive Generative Pre-Trained Transformer and Self-Supervised Learning. *Mathematics* **2023**, *11*. doi:10.3390/math11112451.
22. Mazzeo, R. Editorial. *Topics in Current Chemistry* **2016**, *375*, 1. doi:10.1007/s41061-016-0088-1.
23. Liu, Y.; Lin, H.W. Construction of Interpretation and Presentation System of Cultural Heritage Site: An Analysis of the Old City, Zuoying. *Heritage* **2021**, *4*, 316–332. doi:10.3390/heritage4010020.
24. Platia, N.; Chatzidakis, M.; Doerr, C.; Charami, L.; Bekiari, C.; Melessanaki, K.; Hatzigiannakis, K.; Pouli, P. "POLYGNOSIS": the development of a thesaurus in an Educational Web Platform on optical and laser-based investigation methods for cultural heritage analysis and diagnosis. *Heritage Science* **2017**, *5*, 50. doi:10.1186/s40494-017-0163-0.
25. Pressey, A.; Houghton, D.; Istanbuluoglu, D. The problematic use of smartphones in public: the development and validation of a measure of smartphone "zombie" behaviour. *Information Technology & People* **2023**, ahead-of-print. doi:10.1108/ITP-06-2022-0472.
26. Appel, M.; Krisch, N.; Stein, J.P.; Weber, S. Smartphone zombies! Pedestrians' distracted walking as a function of their fear of missing out. *Journal of Environmental Psychology* **2019**, *63*, 130–133. doi:https://doi.org/10.1016/j.jenvp.2019.04.003.
27. Zhuang, Y.; Fang, Z. Smartphone Zombie Context Awareness at Crossroads: A Multi-Source Information Fusion Approach. *IEEE Access* **2020**, *8*, 101963–101977. doi:10.1109/ACCESS.2020.2998129.
28. Min, B.S. Smartphone Addiction of Adolescents, Not a Smart Choice. *jkms* **2017**, *32*, 1563–1564, [http://www.e-sciencecentral.org/articles/?scid=1023456]. doi:10.3346/jkms.2017.32.10.1563.
29. Huh, J.; Park, S.; Lee, J.E.; Ye, J.C. Improving Medical Speech-to-Text Accuracy with Vision-Language Pre-training Model. 2023, [arXiv:eess.AS/2303.00091].

30. Wahyutama, A.B.; Hwang, M. Performance Comparison of Open Speech-To-Text Engines using Sentence Transformer Similarity Check with the Korean Language by Foreigners. 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 2022, pp. 97–101. doi:10.1109/IAICT55358.2022.9887500.
31. Park, C.; Seo, J.; Lee, S.; Lee, C.; Moon, H.; Eo, S.; Lim, H. BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text. Proceedings of the 8th Workshop on Asian Translation (WAT2021); Association for Computational Linguistics: Online, 2021; pp. 106–116. doi:10.18653/v1/2021.wat-1.10.
32. Saha, S.; Asaduzzaman. Development of a Bangla Speech to Text Conversion System Using Deep Learning. 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2021, pp. 1–7. doi:10.1109/ICIEVicIVPR52578.2021.9564209.
33. Miller, C.; Tzoukermann, E.; Doyon, J.; Mallard, E. Corpus Creation and Evaluation for Speech-to-Text and Speech Translation. Proceedings of Machine Translation Summit XVIII: Users and Providers Track; Association for Machine Translation in the Americas: Virtual, 2021; pp. 44–53.
34. Elakkiya, A.; Surya, K.J.; Venkatesh, K.; Aakash, S. Implementation of Speech to Text Conversion Using Hidden Markov Model. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 2022, pp. 359–363. doi:10.1109/ICECA55336.2022.10009602.
35. Nagdewani, S.; Jain, A. A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION. 2020.
36. Tzoukermann, E.; Van Guilder, S.; Doyon, J.; Harke, E. Speech-to-Text and Evaluation of Multiple Machine Translation Systems. Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track); Association for Machine Translation in the Americas: Orlando, USA, 2022; pp. 465–472.
37. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. 2022, [arXiv:eess.AS/2212.04356].
38. Vrettakis, E.; Kourtis, V.; Katifori, A.; Karvounis, M.; Lougiakis, C.; Ioannidis, Y. Narralive – Creating and experiencing mobile digital storytelling in cultural heritage. *Digital Applications in Archaeology and Cultural Heritage* **2019**, *15*, e00114. doi:https://doi.org/10.1016/j.daach.2019.e00114.
39. Katifori, A.; Roussou, M.; Perry, S.; Drettakis, G.; Vizcay, S.; Philip, J. The EMOTIVE Project-Emotive Virtual Cultural Experiences through Personalized Storytelling. *Cira@ euromed*, 2018, pp. 11–20.
40. Economou, M.; Young, H.; Sosnowska, E. Evaluating emotional engagement in digital stories for interpreting the past. The case of the Hunterian Museum’s Antonine Wall EMOTIVE experiences. 2018, pp. 1–8. doi:10.1109/DigitalHeritage.2018.8810043.
41. Trichopoulos, G.; Aliprantis, J.; Konstantakis, M.; Michalakakis, K.; Mylonas, P.; Voutos, Y.; Caridakis, G. Augmented and personalized digital narratives for Cultural Heritage under a tangible interface. 2021, pp. 1–5. doi:10.1109/SMAP53521.2021.9610815.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.