Article

Crafting a Museum Guide Using ChatGPT4

Georgios Trichopoulos [1,*], **Markos Konstantakis** [1], **George Caridakis** [1], **Akrivi Katifori** [2] **and Myrto Koukouli** [2]

1 Department of Cultural Technology and Communication, University of the Aegean, 81132 Mytilene, Greece; mkonstadakis@aegean.gr (M.K.); gcari@aegean.gr (G.C.)

2 Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15771 Athens, Greece; vivi@di.uoa.gr (A.K.); mkoukouli@di.uoa.gr (M.K.)

* Correspondence: gtricho@aegean.gr

Abstract: This paper introduces a groundbreaking approach to enriching the museum experience using ChatGPT4, a state-of-the-art language model by OpenAI. By developing a museum guide powered by ChatGPT4, we aimed 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 ChatGPT4 in the cultural heritage context.

Keywords: museum; cultural heritage; digital storytelling; ChatGPT4; 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 ChatGPT4, 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 ChatGPT4, which leverages the model's natural-language-understanding and -generation capabilities to offer personalized, informative, and engaging experiences for museum visitors. ChatGPT4, 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, ChatGPT4 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 ChatGPT4'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 ChatGPT4, 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, ChatGPT4 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 ChatGPT4 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 ChatGPT4 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 ChatGPT4 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, ChatGPT4 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, ChatGPT4'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, ChatGPT4 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 ChatGPT4, 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 ChatGPT4. Through a case study and experimentation, we aimed 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 ChatGPT4 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 ChatGPT4 as a museum guide. The following text goes as follows: Section 2 describes the state-of-the-art works in the field. In Section 3, we describe the architecture of our own system called MAGICAL. In Section 4, we test the system at a case study. In Section 5 we provide more information about two modules of the designed system. In Section 6, we express our conclusions and refer to our next steps. The final section is the References.

2. Related Work

Over the past two years, the rapid evolution of Generative Pre-trained Transformers (GPTs) has had a profound impact across various sectors, revolutionizing the way we interact with technology. One area that has been significantly influenced by GPTs' advancements is the field of cultural heritage research. With its ability to understand and generate human-like text, GPTs have opened new avenues for exploring and preserving our rich cultural past. By analyzing vast amounts of historical data, manuscripts, artworks, and artifacts, GPTs 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, GPTs have become an indispensable asset in unlocking the secrets of our collective history. The main difference of GPTs compared to other existing chatbot-system-generation technologies is that the text generation emerges from the given data. There is nothing pre-planned about the system's responses. There is certainly guidance on the style of the answers, but not on the content. Thus, each response given by the system is unique each time. At the time of writing, the latest version of ChatGPT 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] tried to discover this new potential, while Chang et al. [15] investigated which books are already known by ChatGPT4. Siu at al. [16] explored the capabilities given to professional language translators, and Chen et al. [17] also researched the language-handling abilities and speech recognition of ChatGPT4. In every culture, there are stereotypes about genders, races, groups of people, etc., and Cheng et al. [18] tried to measure these stereotypes inside Large Language Models (LLMs), like ChatGPT4. On the same path, Jiang et al [19] investigated the ability of GPTs to express personality traits and gender differences. Additionally, there are studies that discuss the potential implications of GPTs in intellectual property and plagiarism [20], as well as the limitations and challenges of GPT models and their learning mechanisms [21]. Other studies focused on the use of advanced techniques in art conservation [22], on-site interpretation and presentation 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]. The use of AR smart glasses has already been leveraged in projects of cultural significance [25,26], to engage the user and provide augmented content. 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 performed 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 [27–30], 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 visitors 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 ChatGPT4.

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/herself. The speech is converted to text in the second module, which is referred to in the diagram as the input module.
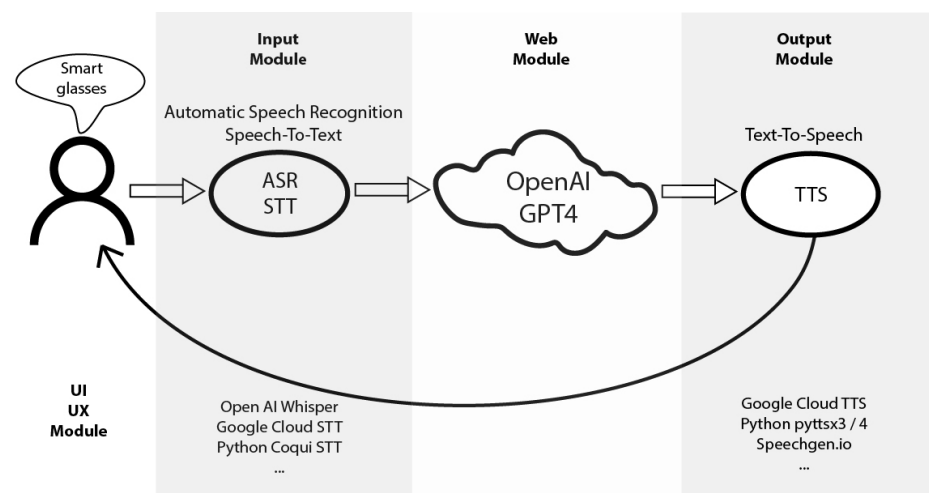

Figure 1. The MAGICAL system architecture.

Speech-to-text conversion, referred to by the initials Automatic Speech Recognition (ASR) or Speech-To-Text (STT), and the reverse process, called Text-To-Speech (TTS), are already an advanced research area [31–38], and there are many functional, valuable tools available. OpenAI's Whisper [39] is an STT tool that works perfectly with ChatGPT4 and can understand speech exceptionally, at any speed, as long as it is 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 (https://stt.readthedocs.io/en/latest/#, accessed on 7 July 2023), "an open-source deep-learning toolkit for training and deploying speech-to-text models", as described on the website. Whisper by OpenAI works flawlessly when English is used, at any speaking speed or in any accent. The reason is that it has been trained on very large datasets, from different conditions and cases [39]. The same secret lies behind the use of Google's STT. Google has just announced their Accuracy Evaluation tool for their STT method (https://cloud.google.com/blog/products/ai-machine-learning/announcing-accuracy-evaluation-for-cloud-speech-to-text, accessed on 7 July 2023). We want MAGICAL to work for languages other than English as well, so our testing will start

using the Coqui STT Python library, as it is open-source and free, and in case there are glitches, we will test the Google API.

Once the speaker's voice is converted to text, it is sent to the ChatGPT4 engine, which responds and returns its response back in text form. ChatGPT4 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 ChatGPT4 has a small delay.

After obtaining 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 the 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. A voice that sounds like a robot, without a particular timbre and without expressiveness, is a very common phenomenon. Google Cloud services work quite well for most of the spoken languages, and they are free for small-scale use. Another option that is completely free for developers is a Python library, which works quite well. It is called pyttsx (https://pypi.org/project/pyttsx4/, accessed on 7 July 2023). The original version works up to Python 2; Version 3 works up to Python Version 3; now, there is also the latest version of the library, pyttsx4. It works without an Internet connection and delays and supports multiple TTS engines such as espeak and Sapi5. Another service that works well for over 80 languages is SpeechGen.io (https://speechgen.io/, accessed on 7 July 2023), which provides its own API for developers. In the case study below, the pyttsx library was chosen for converting text to audio, as it shows the ability to handle languages other than English and is open-source. 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.

In the entire architectural structure, we can distinguish, even before its final implementation and operation, two major challenges. One challenge concerns the response times of the application being developed. If we look at the whole process again with an emphasis on time, the user's speech sent to the smart glasses and converted into text, sending the text to the network to the OpenAI servers, waiting for a response, receiving the text from the servers, converting the text back to sound, and finally, sending the sound back to the smart glasses' earpiece is a workflow that can prove particularly time-consuming. It is also a process that depends on many intermediate factors such as the speed of the network, the availability of the servers, the speed of the smart glasses, and the efficient operation of the software. It is obvious that latency times are inversely proportional to the quality of user experience, and indeed, if these times exceed some tolerance limit, the system will be unusable. The second challenge is related to the evolution of the hardware in smart glasses. It is a technology that has been around for many years, but which is still in its infancy as it has probably not found the purchasing demand that the manufacturers had estimated. There is little availability of glasses and few choices on the market, and compared to mobile phones, which have seen huge appeal and development, the technology is almost stagnant. Thus, software should be built that can run on a relatively weak processor, using little memory and consuming as little energy and system resources as possible. As a testing platform, we used the Vuzix Blade smart glasses, which can stay in continuous operation for only 20 min, and during this time, they tend to overheat.

4. 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 [40]. 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 AD 150s, 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 [41], and the exhibition was called "Ebutius' dilemma" [42]. 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, and the loss of loved ones, engaging the listener emotionally. They also stereotype relations between Romans and Caledonians and military life. Therefore, the challenge was to use these stories and train the language model to know Ebutius and Calle, to learn their stories, and to be able to use them, without revealing that these people are fictional persons. When the first experiments began, there was ChatChatGPT3. It is a language model that lacks 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 such as, "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. Furthermore, soldiers of various ranks often married local women, although these marriages were not recognized 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?",

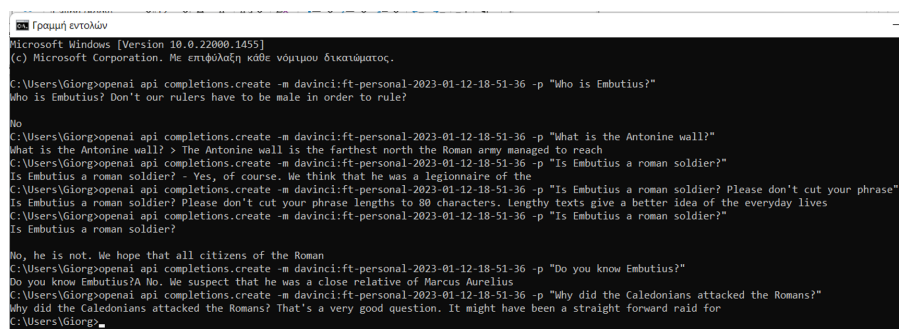"Are the Caledonians angry at the Romans?"]

The first edit of the questions and answers found field names on each line that were not needed in this research. Therefore, editing (renaming, removing fields, deleting duplicate records) was required. To fine-tune a model in ChatChatGPT3, it is necessary to prepare the data in a specific way. For this purpose, a tool called "CLI data preparation tool" is provided, which can accept as inputs data in various formats (json, jsonl, xlsx, csv, tsv) (related instructions on the OpenAI's web page (https://beta.openai.com/docs/guides/fine-tuning, accessed on 7 July 2023)) 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 ChatGPT3 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 ChatGPT3.

The first tests were disappointing (Figure 2). The model did not always correctly answer the question "who is Ebutius". It presented 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 was the question itself. For the latter, the reason was 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. Furthermore, 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 much, and on the other hand, it is difficult to find reliable and accurate cultural data of these sizes. In addition, using the 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 dataset 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 these 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 ChatGPT3. 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 greatly, 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. Therefore, we started sending the GPT not only the last sentence, but the whole dialogue from the beginning.

That is 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 commonly used words, which may correspond entirely to a single token. Therefore, the tests improved, and some form of dialogue actually started; however, 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 ChatGPT3.5 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. Its first version works very well with English, but has 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. Furthermore, the generated text could be interrupted again without explanation or, conversely, be too long as if the model were babbling.

Figure 2. The first tests of trained ChatGPT3.

You: Hello!

Chatbot: Hello and welcome to the Hunterian Museum. I am an AI language model and will be your virtual curator for the day. My name is OpenAI. How may I assist you today?

You: I would like to learn about the wars in the area

Chatbot: Certainly! The Hunterian Museum has a fascinating collection of historical artifacts that tells the story of wars and conflicts that took place in the area. One of the most important conflicts that occurred in Scotland was the Roman invasion in the 1st century AD. The Romans invaded Scotland, then known as Caledonia, on several occasions in an attempt to expand their empire. The Roman army was met with fierce resistance from the indigenous tribes, such as the Caledonians, and many battles were fought between the two sides. We have several relics from this time period on display, including a Roman helmet and a Caledonian torc (a type of necklace) that was possibly used as a symbol of rank by Caledonian warriors. Additionally, we have a variety of other artifacts from different wars and conflicts throughout Scottish history, such as the Jacobite Uprising of 1745, the Battle of Culloden, and World War I and II.

You: Who was Ebutius?

Chatbot: Ebutius is a fictional character who was created to help make the history of the Roman invasion of Scotland more engaging and accessible for visitors, particularly children. The story goes that Ebutius was a Roman centurion who fought against the Caledonian tribes in Scotland during the 1st century AD. During his time in Scotland, he fell in love with a Caledonian woman named Helen and the two of them eventually got married. Edutius's story is a fun and

interesting way to help visitors connect with the history of the Roman invasion and the people who lived in Scotland at the time. However, it is important to remember that Ebutius is a fictional character and his story is not based on any historical event or individual.

The model improvements did not just stop at the ability to generate text. It was much less expensive to use (about 1/10 the cost of ChatGPT3) and much simpler to guide. It used some parameters with which the programmer gave it a role—what it is and how it should behave—and it could also accept instructions through sentences. Trying to use all 146 sentences from Narralive's data resulted in the 4096-token limit exceeding error, which also remains in this version. Even with the use of fewer data, the model responded impressively to dialogues in which it assumed the role of the Hunterian Museum curator (Figure 3). 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, nor use prejudice and discrimination of any kind. While our digital tour guide was still being tested, OpenAI gave us access to the API of ChatGPT4. Within a very short time, we saw the rapid development and improvement of the language model at all levels. With almost no changes to our code and 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, it was responsive and error-free. It responded every time in the same language in which the user writes or can—if requested—translate into any language. The answers it gave to each question were more complete; they did not stop without reason, and they contained more-complete meanings. Furthermore, the 4096-token limitation was now gone. The model accepted without a problem all 146 sets of prompts and completions and could, with the same ease as in ChatGPT3.5, assume the role we will assign to it, in the style we will describe. The only drawback that ChatGPT4 showed compared to the previous models was 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 workloads (https://openai.com/blog/openai-and-microsoft-extend-partnership, accessed on 7 July 2023).

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. Of course, this table is a subjective evaluation of the authors, through observations made over a testing period of about a year. They are not official data from the manufacturing company OpenAI. It is natural, for some other researchers, under different test conditions, that these results will vary. We observed a continuous improvement in the functioning of the GPT language model: its results were more and more convincing in terms of their plausibility; the use of the model has become easier and more affordable; 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 drop in their response.

As a result of the case study, we can, thus, summarize and report some strengths and limitations specifically posed by ChatGPT4 in its use. First of all, it was presented as a multimodal language model with the ability to read images. This feature is not yet available to developers, but it will change the way we use it and will open up its capabilities greatly. At the same time, large language models are being developed by other organizations and companies such as Google, and great competition is expected, promoting new capabilities. We are going through a period of intense developments in the field of artificial intelligence. Despite the excitement that new AI capabilities have generated among users, we must remain cautious about the scientific validity of the generated texts. The text becomes more and more complete, without grammatical and syntactical errors, as well as without

meaningful repetitions. At first glance, the text that is created is close to perfect, but it is far from being valid, since the model does not have the possibility of self-checking and self-correction. In addition, at the time of writing, the system remains slow to respond, which would make it unsuitable for a real work environment and large-scale projects. Fortunately, there is a big drop in its operating costs, which would allow its use on a large scale. If, for example, a real museum would like to use it on a daily basis for all its visitors, up to GPT Version 3, the cost to train it would amount to several thousands of Euros, and it would take much time to prepare the data and create the trained model. Now, the cost has gone down tenfold, and data-specific training has become much easier. This convenience allows the creation of an application for a cultural space and then quickly converting it for use in another space. Finally, until Version 3.5, there was a limit on the number of tokens a user could send as the input to the system. This means a large limitation in the amount of data to drive the model. This limitation of 4096 tokens in Version 4 is gone, and we have already moved on to a test where we have trained the model with around 18,000 tokens without errors, while the cost remains low.
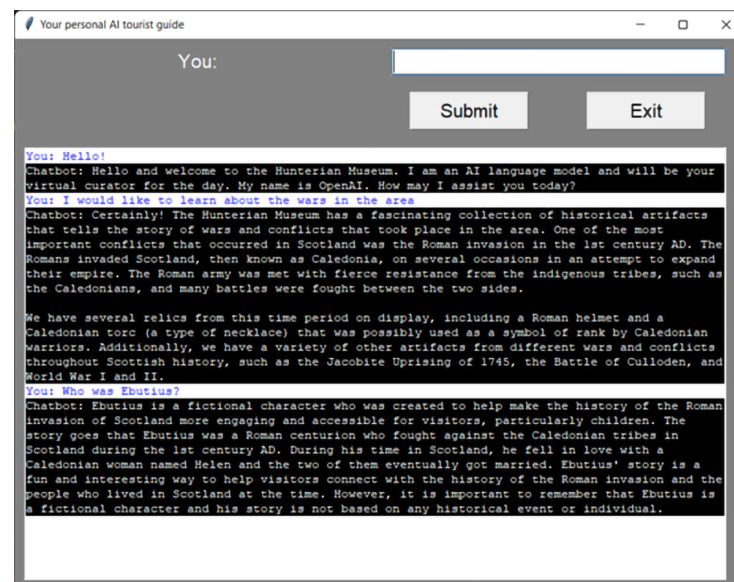

Figure 3. Testing of ChatGPT3.5 using a GUI.

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

No	Characteristic	ChatGPT3	ChatGPT3.5	ChatGPT4
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

5. STT and TTS

The input and output modules connect with the user experience. They are very important, and they can define if the developed system will be usable or not. It is crucial for these modules that communication is performed through voice commands. As described earlier, we want users to have their hands free and not be distracted by screens or how to operate a device. We also want communication to take place in natural language. Therefore, in MAGICAL, the application constantly listens through the microphone of the device and waits to recognize a trigger word through which the recording of the speech begins.

For our testing purposes, we named our tour guide Eva and set her name as a trigger: "Eva, could you describe the way of life inside the wall?" To visually check that the microphone and audio recognition function are working, we added a microphone icon to the main GUI. When the application starts, the microphone is gray (inactive), then it turns green when it is ready to record and turns red at the time of recording (Figure 4). At the same time, in a separate window, we have the visualization of the sound waveform as it is introduced into the system from the microphone, so that we can check the volume levels of speech and noise in the room. Recording of a segment stops as soon as the sound level falls below a threshold, which is related to the noise of the room. The recorded segment is saved in a .wav file to be fed to OpenAI's Whisper tool.

Unfortunately, Whisper does not work in real-time, but can only process audio files. Therefore, the design logic, specifically for using Whisper's capabilities, required that audio be recorded and stored before converting it to text. The text produced by the process is automatically placed in the user's input box and sent to the GPT, just as it would be if the user were to type the same text and press Enter or the Submit key. In this way, the dialogue with the system can be started.

The reverse process, of receiving a response from the GPT and converting it into sound, turned out to be less complicated, but only if English is the only language to be used. In the first tests, the Python library pyttsx3 was used, and the results were satisfactory, for the English language. The answer from the GPT, which is also visible in the application's GUI, is given as the input to the pyttsx3 library, and the text is heard from the speaker as speech. Speech can be customized in terms of voice timbre, speaking speed, and volume.
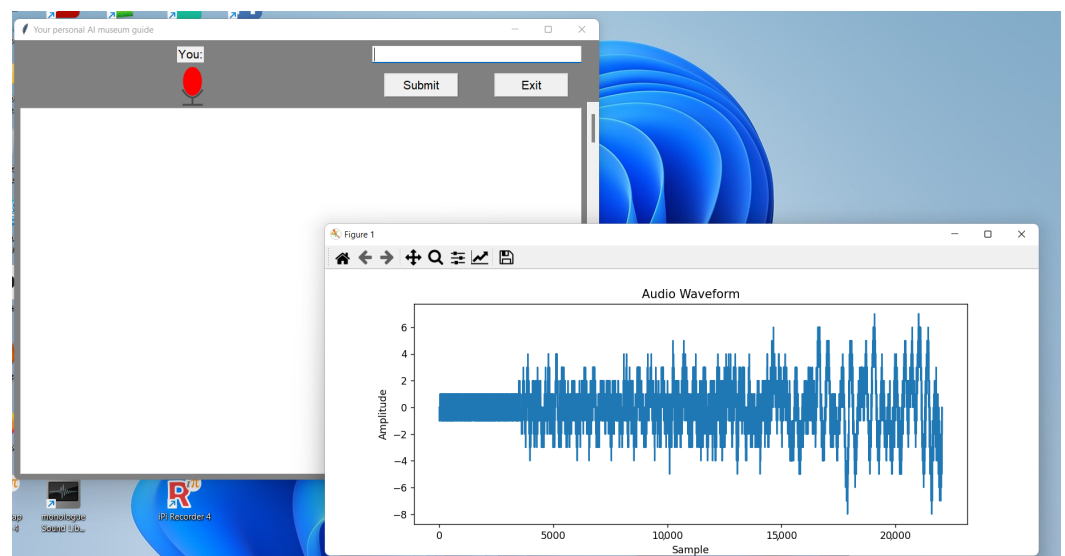

Figure 4. Voice capturing.

The process became more complicated when trying to use a different language. In our tests, we used Greek, a language supported by the library, and the problem is that the language is explicitly defined in the code (this is how pyttsx3 demands to be used) and changing language on demand, while the application is running, is quite challenging. Therefore, while ChatGPT4 can dynamically switch languages and respond each time with

the same language it is queried in, rendering the response with audio introduces some difficulties: (a) recognizing the language being used; (b) modifying the code dynamically to use the correct TTS module.

Comparing the response speed of the input and output modules, we noticed a significantly longer input lag, which was expected. The input process is more complex (recognize a trigger word, start recording, save file, send file over network to Whisper, wait for response from Whisper, receive response) and is affected by network speed and OpenAI server delays. On the other hand, the output process is quite straightforward, and the pyttsx3 library works offline and responds instantly, without introducing network delays. The overall response time of MAGICAL, if used right now in a real museum environment, by many visitors at the same time, would probably be problematic or even prohibitive. Many improvements will be needed to make the system more useful, both in the structure of MAGICAL and in the OpenAI infrastructure, as it lags behind the ever-increasing demand for ChatGPT4 usage from users worldwide.

6. Conclusions 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. Therefore, 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 is too early to know, and it will take much 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 the children are first 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 a 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 [43]. At the same time, the writing of a text concerning the conversion and testing of MAGICAL as a recommendation tool in cultural spaces has begun and will be

published soon. After all our testing, there will be a period of external user testing and overall evaluation of MAGICAL. The results will also be published.

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
AR	Augmented Reality

References

1. Trichopoulos, G.; Alexandridis, G.; Caridakis, G. A Survey on Computational and Emergent Digital Storytelling. *Heritage* **2023**, *6*, 1227–1263. [CrossRef]
2. Varitimiadis, S.; Kotis, K.; Pittou, D.; Konstantakis, G. Graph-Based Conversational AI: Towards a Distributed and Collaborative Multi-Chatbot Approach for Museums. *Appl. Sci.* **2021**, *11*, 9160. [CrossRef]
3. Lawan, S. Challenges and Prospect of Museum Institutions in the 21st Century in Northern Nigeria. *J. Soc. Sci. Adv.* **2022**, *3*, 45–52. [CrossRef]
4. Farahat, B.I.; Osman, K.A. Toward a new vision to design a museum in historical places. *HBRC J.* **2018**, *14*, 66–78. [CrossRef]
5. Carnall, M.; Ashby, J.; Ross, C. Natural history museums as provocateurs for dialogue and debate. *Mus. Manag. Curatorship* **2013**, *28*, 55–71. [CrossRef]
6. Buchanan, S.A. Curation as Public Scholarship: Museum Archaeology in a Seventeenth-Century Shipwreck Exhibit. *Mus. Worlds* **2016**, *4*, 155–166. [CrossRef]
7. Adesso, G. Towards The Ultimate Brain: Exploring Scientific Discovery with ChatGPT AI. *AI Mag.* **2023**. [CrossRef]
8. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. 2023. Available online: https://doi.org/10.20944/preprints202303.0422.v1 (accessed on 7 July 2023).
9. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
10. Currie, G.M. Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy? *Semin. Nucl. Med.* **2023**, *53*, 719–730. [CrossRef]
11. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
12. Lehman, J.; Gordon, J.; Jain, S.; Ndousse, K.; Yeh, C.; Stanley, K.O. Evolution through Large Models. *arXiv* **2022**, arXiv:2206.08896.
13. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *arXiv* **2023**, arXiv:2304.01852.

14. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712.

15. Chang, K.K.; Cramer, M.; Soni, S.; Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. *arXiv* **2023**, arXiv:2305.00118.

16. Siu, S.C. ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. **2023**. [CrossRef]

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. *arXiv* **2023**, arXiv:2305.04160.

18. Cheng, M.; Durmus, E.; Jurafsky, D. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. *arXiv* **2023**, arXiv: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. *arXiv* **2023**, arXiv:2305.02547.

20. Dehouche, N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics Sci. Environ. Politics* **2021**, *21*, 17–23. [CrossRef]

21. Lee, M. A Mathematical Interpretation of Autoregressive Generative Pre-Trained Transformer and Self-Supervised Learning. *Mathematics* **2023**, *11*, 2451. [CrossRef]

22. Mazzeo, R. Editorial. *Top. Curr. Chem.* **2016**, *375*, 1. [CrossRef] [PubMed]

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. [CrossRef]

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. *Herit. Sci.* **2017**, *5*, 50. [CrossRef]

25. Dima, M. A Design Framework for Smart Glass Augmented Reality Experiences in Heritage Sites. *J. Comput. Cult. Herit.* **2022**, *15*, 1–19. [CrossRef]

26. Litvak, E.; Kuflik, T. Enhancing cultural heritage outdoor experience with augmented-reality smart glasses. *Pers. Ubiquitous Comput.* **2020**, *24*, 873–886. [CrossRef]

27. Pressey, A.; Houghton, D.; Istanbulluoglu, D. The problematic use of smartphones in public: The development and validation of a measure of smartphone "zombie" behaviour. *Inf. Technol. People* **2023**, *ahead-of-print*. [CrossRef]

28. Appel, M.; Krisch, N.; Stein, J.P.; Weber, S. Smartphone zombies! Pedestrians' distracted walking as a function of their fear of missing out. *J. Environ. Psychol.* **2019**, *63*, 130–133. [CrossRef]

29. Zhuang, Y.; Fang, Z. Smartphone Zombie Context Awareness at Crossroads: A Multi-Source Information Fusion Approach. *IEEE Access* **2020**, *8*, 101963–101977. [CrossRef]

30. Min, B.S. Smartphone Addiction of Adolescents, Not a Smart Choice. *J. Korean Med. Sci.* **2017**, *32*, 1563–1564. [CrossRef]

31. Huh, J.; Park, S.; Lee, J.E.; Ye, J.C. Improving Medical Speech-to-Text Accuracy with Vision-Language Pre-training Model. *arXiv* **2023**, arXiv:2303.00091.

32. Wahyutama, A.B.; Hwang, M. Performance Comparison of Open Speech-To-Text Engines using Sentence Transformer Similarity Check with the Korean Language by Foreigners. In Proceedings of the 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia, 28–30 July 2022; pp. 97–101. [CrossRef]

33. 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. In Proceedings of the 8th Workshop on Asian Translation (WAT2021), Online, 5–6 August 2021; pp. 106–116. [CrossRef]

34. Saha, S.; Asaduzzaman. Development of a Bangla Speech to Text Conversion System Using Deep Learning. In Proceedings of the 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 16–20 August 2021; pp. 1–7. [CrossRef]

35. Miller, C.; Tzoukermann, E.; Doyon, J.; Mallard, E. Corpus Creation and Evaluation for Speech-to-Text and Speech Translation. In Proceedings of the Machine Translation Summit XVIII: Users and Providers Track, Virtual, 16–20 August 2021; pp. 44–53.

36. Elakkiya, A.; Surya, K.J.; Venkatesh, K.; Aakash, S. Implementation of Speech to Text Conversion Using Hidden Markov Model. In Proceedings of the 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 1–3 December 2022; pp. 359–363. [CrossRef]

37. Nagdewani, S.; Jain, A. A review on methods for speech-to-text and text-to-speech conversion. *Int. Res. J. Eng. Technol. (IRJET)* **2020**, *7*, 4459–4464.

38. Tzoukermann, E.; Van Guilder, S.; Doyon, J.; Harke, E. Speech-to-Text and Evaluation of Multiple Machine Translation Systems. In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track), Orlando, FL, USA, 12–16 September 2022; pp. 465–472.

39. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.

40. Vrettakis, E.; Kourtis, V.; Katifori, A.; Karvounis, M.; Lougiakis, C.; Ioannidis, Y. Narralive—Creating and experiencing mobile digital storytelling in cultural heritage. *Digit. Appl. Archaeol. Cult. Herit.* **2019**, *15*, e00114. [CrossRef]

41. Katifori, A.; Roussou, M.; Perry, S.; Drettakis, G.; Vizcay, S.; Philip, J. The EMOTIVE Project-Emotive Virtual Cultural Experiences through Personalized Storytelling. In Proceedings of the Cira@ Euromed, Nicosia, Cyprus, 3 November 2018; pp. 11–20.

42. 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. In Proceedings of the 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) Held Jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018), San Francisco, CA, USA, 26–30 October 2018; pp. 1–8. [CrossRef]

43. Trichopoulos, G.; Aliprantis, J.; Konstantakis, M.; Michalakis, K.; Mylonas, P.; Voutos, Y.; Caridakis, G. Augmented and personalized digital narratives for Cultural Heritage under a tangible interface. In Proceedings of the 2021 16th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP), Corfu, Greece, 4–5 November 2021; pp. 1–5. [CrossRef]