A Framework for Creating Customized Virtual Environments

by

Riccardo Caprile

DIBRIS, Università di Genova

Via Opera Pia, 13 16145 Genova, Italy   
http://www.dibris.unige.it/

A drawing of a person

Description automatically generated

Laurea Magistrale in Computer Science

MSc in Computer Science

Data Science and Engineering Curriculum

**A Framework for Creating Customized Virtual Environments**

**Riccardo Caprile**

Advisor: Manuela Chessa Examiner: Nicoletta Noceti

January, 2023

**Abstract**

This thesis aims to exploit generative language models to create customized virtual reality environments.  
In many fields of application, such as exergames for cognitive or physical training, one of the main issues is to create a wide range of simulated situations, thus avoiding the onset of boredom and habituation phenomena. However, creating many simulated situations may be a long and tedious task.   
This thesis aims to cope with this issue, creating a framework for generating different virtual reality scenarios described in natural language. The framework will allow unskilled people to add virtual scenarios and simulations to existing sw, such as cognitive exergames.  
The student will analyse the use of generative language models (e.g. chatgpt) and their use to create virtual environments (e.g., in Unity).  
A further development will be the combination with real-world elements, detected in real-time, to create dynamic environments where virtual and real objects coexist.

Table of Contents

[Chapter 1 Introduction 6](#_Toc161680908)

[1.1 Motivations and Goals 6](#_Toc161680909)

[1.2 Virtual Reality (VR) 7](#_Toc161680910)

[1.2.1 Applications 9](#_Toc161680911)

[Chapter 2 Large Language Model 12](#_Toc161680912)

[2.1 Definition 12](#_Toc161680913)

[2.2 Use Cases 13](#_Toc161680914)

[2.3 GPT 14](#_Toc161680915)

[2.3.1 ChatGPT 15](#_Toc161680916)

[2.4 LLM and VR 15](#_Toc161680917)

[Chapter 3 Methods 16](#_Toc161680918)

[3.1 Software 16](#_Toc161680919)

[3.1.1 Unity 16](#_Toc161680920)

[3.1.2 API OpenAI 16](#_Toc161680921)

[3.1.3 Roslyn C# Runtime Compiler 16](#_Toc161680922)

[3.1.4 Vocal Commands 16](#_Toc161680923)

[3.2 Hardware 16](#_Toc161680924)

[3.2.1 Meta Quest 2 16](#_Toc161680925)

[Chapter 4 Design of the System 17](#_Toc161680926)

[Chapter 5 Implementation 18](#_Toc161680927)

[Chapter 6 Results 19](#_Toc161680928)

[6.1 System Testing 19](#_Toc161680929)

[6.1.1 Experiments 19](#_Toc161680930)

[6.1.2 Script Generation: Execution Time 19](#_Toc161680931)

[6.1.3 Script Generation: Error Rate 19](#_Toc161680932)

[6.2 User Testing 19](#_Toc161680933)

[6.2.1 Testers Description 19](#_Toc161680934)

[6.2.2 System Usability Scale (SUS) 19](#_Toc161680935)

[6.2.3 User Experience Questionnaire (UEQ) 19](#_Toc161680936)

[Chapter 7 Conclusions 20](#_Toc161680937)

[7.1 Future Works 20](#_Toc161680938)

[Bibliography 21](#_Toc161680939)

# Introduction

The first chapter briefly gives a summary of the topics covered inside the other chapters and the principal objectives of this thesis and an introduction to Virtual Reality (VR).

## Motivations and Goals

Large language models trained for code generation can be applied to speaking virtual world into existence, the main objective is using them for creating custom virtual environments suitable with user requests. In the recent years, AI models have reached a very high level of reliability when a question has asked them. So, in this thesis we wanted to exploit those models and to give to the user a framework that can overcome the problem of having different environments for different situations with a small effort. We want to give the possibility to build environments with the support of the natural language in order to have a framework accessible to unskilled people not related to computer science world. For instance, a physiotherapist would ask for a virtual environment that is tuned the needs of a patient with an unusual injury. Students in a laboratory could inspect, analyse, and become familiar with materials or machines that they could never handle, because of the school’s budget or because they are too dangerous to be used inside such structures. There are thousands of possible examples of custom virtual environments.

Without the support of a framework like this, the building of the environment would a boring a tedious task, that must be repeated hundreds of times.

Firstly, we wanted to define and explain the Large Language Models and what is the relationship between them and Virtual Reality, because the framework is based on them, and the combination of these tools will produce the desired result. Without LLM we could not have the outputs requested for creating the customized world.

Following, we wanted to describe hardware and, especially, software needed for the realization of the framework, giving technical aspects of those methods. Then, we would analyse the design of the system and understand how the tools described before are combined with.

At this point, we must go deeper into the design and the implementation of the system, analysing all the C# scripts, and how every step of the design has been implemented.

In the last part of the thesis, we have a section for the results obtained by the framework. The “users analysis” is going to be based on their feedback with a couple of questionnaires, in addition to this there is going to be a technical analysis based on how the system worked in different situations (eg. Error rate and execution time).

To conclude this work, we would like to give how the framework could be improved in future works.

## Virtual Reality (VR)

Virtual Reality (VR) is a technology which allows a user to interact with a computer-simulated environment that can be a representation of the real world or an imaginary world. The main objective is to give an immersive experience of a virtual world. The environment is typically accessed via a display, usually a wearable display (eg. Head mounted display (HMD), gloves or body suits).

The illusion of “being there” is affected by motion sensors that pick up the user’s movements and adjust the view on the screen accordingly in real time.

VR can be divided into five categories:

* **Non-Immersive:** It refers to a virtual experience through a computer where you can control some characters within the software, but the environment is not directly interacting with you. Basically, you are dealing with a virtual world, but you are not at the centre of attention in the game.

All basic forms of gaming devices such as: Playstation, Xbox, Computers, etc, provide us with a non-immersive virtual reality experience.

* **Immersive:** This type of experience is what most people think what they come across with Virtual Reality. It is based on wearable displays to track user’s movements and present the VR information depending on the position of users, which allows them to experience 360 degrees of the virtual environment. It feels

like you are within the virtual world physically.

* **Semi-Immersive:** As the name suggests, it is a mix of non-immersive and immersive virtual reality. It is a 3D space or virtual environment where you can move on your own, either through a headset or a computer screen. All the activities are concentrated toward you, but, on the other hand, you have no real physical movements other than your visual experience.

Immagine che contiene schermata, collage, uomo, arte

Descrizione generata automaticamente

1.1 The three categories of Virtual Reality

In between immersive and non-immersive virtual reality, seen before, there is also:

* **Augmented Reality:** It combines the real world and computer-generated content. It is a technology that enhances the real-world environment with digital information and objects, creating a more immersive and interactive experience. Users can access digital information in a more engaging and interactive way.

An example of augmented reality are interactive displays;for instance, when we scan an object in a shop, we could visualize some extra information about that article.

One of the first appearance of Virtual Reality devices similar to Head Mounted displays was in the first half of 1800s with the stereoscope. It was built by following research that demonstrated that the brain combines two photographs of the same reference object taken from different points to make the image appear to have a sense of depth and immersion.

In 1956, the cinematographer Morton Heilig created Sensorama [[1]](#One) the first VR machine. It combined full colour 3D video, audio, vibrations, smell and atmospheric effects.

His second device created was the so called Telesphere Mask which was the first HMD. It provided stereoscopic 3D images with wide vision and stereo sound.

Following this creation, Ivan Sutherland in 1965 presented his concept of the “Ultimate Display”. Basically, his idea was based on a virtual world viewed through an HMD which replicated the reality so the user would not be able to distinguish between real world and virtual world.

*"The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. Handcuffs displayed in such a room would be confining, and a bullet displayed in such a room would be fatal.”* *[[2]](#Two)*

In 1968, he actually created the first virtual reality HMD, called “The Sword of Damocles”. It was connected to a computer and gave only the possibility to show some simple virtual wire shapes. These 3D models changed perspective when the user moved his head thanks to the tracking system.

Talking about more recent years, in 1991 Antonio Medin, a NASA scientist, designed a VR system which the main scope was to drive Mars robot from the Earth in real-time, the name of this system was “Computer Simulation Teleoperation”.

From this point, software house like SEGA and Nintendo, started to develop some Virtual Reality headsets available for the general public to purchase.

In 2012, this revolution exploded thanks to a Kickstarter campaign for the Oculus Rift raising a huge amount of money.

In 2014, with the acquisition of Oculus by Facebook, many companies started the development of their own Headsets.

Sony announced that they were working on Project Morpheus for Playstation 4.

Google released the “Cardboard” – a low-cost stereoscopic viewer designed for smartphones.

Similarly, Samsung announced the “Samsung Gear VR”, a headset that uses the smartphone Samsung Galaxy as a viewer.  
By 2016 hundred of companies started developing VR products: HTC, Google, Apple, Amazon, Sony, Samsung, Microsoft.

In the end, with Oculus was demonstrated that Virtual Reality has incredibly progressed and is being used in a wide range of ways, starting from gaming experiences to teaching new skills and experience brand new virtual journeys.

### Applications

So now, we would like to introduce some of the applications of Virtual Reality

**Education**

The first interesting application for Virtual Reality is the educational scope. Nowadays, is not used so much but there are many promising examples and studies that demonstrate how beneficial VR could be in the educational environment.

Firstly, the usage of VR could increase students’ attention to what is happening inside the VR environment by keeping them always motivated and focussed.

Because of this technology, they are extremely more interested and engaged in what they are learning at school, avoiding any kind of distraction.

VR has also the potential to transport students to different environments, allowing them to explore new places or environments, that are unreachable to humans, safely and efficiently.

**Medical Training**

One of the most interesting and important applications of Virtual Reality is the possibility that is given to professional doctors to practise operations on “virtual” patients, which has brought better outcomes and reduced the number of mistakes during medical operations, with a better preparation the rate of success is higher.

Technically speaking, students can train in an interactive virtual environment that can be programmed with different scenarios and parameters. In this way, they have the opportunity to test different real-life settings.

The simulation could contain a video that shows to students how the operation should be done or how a tool should be used. A lot of different information could be displayed during the test.

It could be also programmed a patient-specific environment, in this case they could hundreds of possible combinations of training.

**Exergaming, Fitness and Sports**

Virtual Reality can also be used for practising sports and keep us fit and in good shape. This has been possible thanks to a genre of games called exergames, in which players participate in physical activities to reach achievements inside the game. The main objective of this type of exercises is to overcome the usual sedentary activity that is at the base of traditional games by using body activity as input and not just by pressing only buttons on keyboard, mouse, or controller.

VR technology can also be used in sports when athletes need to improve their skills, or they have to recover from an injury with the help of therapy and rehabilitation.

**Therapy**

One interesting usage of Virtual Reality is therapy, especially mental therapy. VR can be used by people who suffer of post-traumatic stress disorder, it can be a way to trigger traumatic event but in a safe place at the presence of a therapist in order to receive help when they have to deal with the traumatic event.

In addition to this, when a patient needs a safe and tranquil place where he can calm down, he could use Virtual Reality to create an environment where he can destress and isolate himself from the external world.

**Entertainment**

The last application is the most famous and used one among the general public. Nowadays, Virtual Reality is used only for entertainment scopes such as watching movies and TV shows or play simple video games just for fun or try different experiences that they could have daily with usual computer and monitors.

In conclusion, we can say that Virtual Reality has an infinite number of applications and usage if we consider how many environments we could create with the help of this technology, and that is what we have done: to give to the user the possibility of creating customized virtual environment as simple as possible.

# Large Language Model

Because the framework implemented it is based on a large language model such as ChatGPT, developed by OpenAI, we want to provide, in this chapter, information about these models and their connection with the Virtual Reality development world.

## Definition

A large language model (LLM) is a deep learning algorithm that can perform multiple Natural Language Processing (NLP) tasks, such as: text generation, code generation and information retrieval.

LLMs and Generative IA are obviously linked to each other, because LLMs are a type of generative AI that has been built for the generation of text-based content.

Basically, LLMs are designed to understand and generate text like a human basing their knowledge on the vast amount of data used to train them. They have the ability to infer from context and generate coherent and relevant responses understandable by humans.

These kinds of models are based on a transformer architecture, which is perfect when there is the necessity to handle sequential data like text input.

It was presented to the public in 2017 in a research paper by Google called “Attention is All You Need” [3].

Transformers are neural networks that learn context and understanding through sequential data analysis. They use the technique called “Attention” that helps them to identify how distant data elements influence and depend on one another.

Transformers took inspiration from the encoder-decoder architecture. But the power of this algorithm is that it does not perform data processing in sequential order, but it parallelizes all the operations making the training faster.

Now, we know a little bit more what there is behind a LMM, but how do they work?

As we said before, LMMs are based on a transformer model and works by receiving an input, encoding it with an encoder, and, in the end, decoding it to produce the output desired. But before all these operations, we need to **train** them, so they can fulfil general functions, and **fine-tuning**, which enables them to perform more specific operations.

**Training:** During this step, models are exposed to a wide amount of unlabelled textual datasets taken from websites such as: Wikipedia, GitHub, or others.

The models must learn patterns, structures, meaning of the words, relationships networks and semantic knowledge present in the text corpus. At this point the datasets are processed without specific instructions, unsupervised learning.

**Fine-Tuning:** A language model needs to be trained to perform a specific task or domain.

This operation is possible because we start from the LLM, then we train it on a labelled dataset (Supervised Learning) for the specific task or domain wanted.

Thanks to Fine-Tuning, the performances of the LLM can be improved on the specific task or domain by adjusting the waits of the model to better fit the data.

In this case the labelled data consists of pairs of input and output data.

Of course, the input data is the data that will be given to the LLM, and the output data is the data that the LLM is expected to generate. It is a very efficient and simple algorithm.

## Use Cases

Large language models can be used for several reason:

* Information Retrieval: Whenever you use a search engine and their feature, you are relying on a large language model to provide information in response to a query written by the user.
* Sentiment Analysis: LLMs enable users to analyse the sentiment of textual data.
* Text Generation: LLMs are able to generate text based on inputs. They can produce an example of text when prompted.
* **Code Generation**: It is very similar to Text Generation and this the purpose we are going to exploit late in the framework description. The LLM will generate a C# script, according to our requests, that will be executed and will create the virtual environment asked by the user.
* Chatbots and conversational IA: Large language models enable customer service chatbots to have a conversation with customers, interpret the meaning of their queries or responses, and offer them responses to clarify their doubts.

Furthermore, large language models can be found in a huge number of fields such as Healthcare and Science (understand proteins, molecules and DNA or be a helpful tool in the development of vaccine), Marketing (usage of sentiment analysis for generate aimed marketing campaign), Legal (to translate documents written in basic language to more complex and legalese language).

## GPT-3

GPT stands for Generative Pre-Trained Transformer 3 which is a large language model released by OpenAI [4] in 2020.

It belongs to a family of AI models.

The latest GPT model is GPT-4 which is the fourth generation, although various versions of GPT-3 are still widely used and available. These models are the ones we are going to use for testing our virtual environment examples.

There are many models in the GPT-3 family, some are used for different purposes than others. In the presentation paper [5] published, there is a table where there are mentioned 8 different sizes of the main GPT-3 model:

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Parameters | Number of layers | API Name |
| GPT-3 Small | 125M | 12 | n/a |
| GPT-3 Medium | 350M | 24 | ada |
| GPT-3 Large | 760M | 24 | n/a |
| GPT-3 XL | 1.3B | 24 | babbage |
| GPT-3 2.7B | 2.7B | 32 | n/a |
| GPT-3 6.7B | 6.7B | 32 | curie |
| GPT-3 13B | 13.0B | 40 | n/a |
| GPT-3 175 or GPT-3 | 175.0B | 96 | davinci |

The parameters in GPT-3, like any neural network, are the weight and the biases of the layers. The sizes depend on the number of layers, so more layers a version has more parameters has.

Weights are the most important parameters from which the model learns from the training data. For example, if a model sees the word “Dog” followed by the word “bark”, it will assign a higher weight to this pair of words. Next time, the model will be able to predict more likely the word “bark” after “dog”.

Bias is a parameter used as an adjusting factor to tune the prediction of an entire layer to the more accurate side. Basically, it is used to adjust the output values and influences how easy it is for a node to be activated.

Most of the models listed above are accessible through API, we will se later which models there are available, and which are their purposes.

### ChatGPT

ChatGPT [6] is a chatbot developed and released by OpenAI in November 2022. As the name suggests, it is based on the large language model GPT, described before.

It gives the user the possibility to have a conversation with a desired length, style and language.

ChatGPT is available for everybody in two versions, one built with GPT-3.5 and the other with GPT-4.

This chatbot was trained on a massive corpus of text data, around 570GB of datasets, including software manual pages, web pages, books, and other kind of sources.

The main objective of ChatGPT is to have a humanlike conversation trying to answer to users’ questions as much precisely as possible.

### Prompt Engineering

Prompt engineering is an artificial intelligence engineering technique that comprehends the process of refining LLMs with specific prompts and recommended output, and, at the same time, the process of refining input to get the best output from an LLM.

Prompt engineering is fundamental for creating better AI-powered services and getting better results from AI generative tools.

For example, we will face this problem in the testing chapter where we will see what kind of sentences have been used as input and the expected and unexpected results we have gotten.

There four types of prompts:

* Zero-shot: It is the simplest and most direct type of prompt. It is given to the generative IA a direct instruction or question without adding extra information about it. This type of prompts technique tests the model’s ability to produce relevant output without relying on prior examples (e.g “List me some Computer Science projects that I can develop”)
* One-shot: In this case a single example guides the model’s output. This example can be a question-answer pair or a specific template. This method has the objective to align the model’s output more closely with the user’s desired format of response.
* Few-shot: This technique, similar to One-shot, involves providing the generative IA with some examples to guide it to what you would like it to imitate. This helps the models to understand the user’s requirements better, generating output that closely follow the same pattern of the given examples.
* Chain-of-Thought (CoT): This method is designed to guide the model through a logical reasoning process. The idea is not just asking for an output, but it is a strategy that incorporates demonstrations that tell the model how to arrive at the correct answer step by step. Basically, it encourages the large language model to explain its reasoning. With this kind of prompt structure, we get the solution and understand how the AI reasoned, producing a detailed output.
* Zero-shot Chain-of-Thought (CoT): It is a technique similar to Zero-shot prompting explained earlier, but this approach adds an instruction at the end of the request: “Let’s think step by step”.

The model is going to be able to generate a chain of thoughts from this simple instruction, and of course the result will be more accurate as well.

In addition to this, there are some ideas for refining prompts whenever you get stuck, and most of them were used during the conversations we had with ChatGPT when a certain script had been required.

Firstly, it is recommended to repeat key words and use caps to stress important points or specific details desired. Another idea is to use frequently the verb “must”.

Try to write long requests repeating the same statement many times but in different places of the sentence.

## LLM and VR

In this subchapter, we would like to talk about the relationship between large language models and virtual reality nowadays.

It has been possible to find just one paper [7] about this topic; it means that this relationship has not been explored and exploited so much at the moment.

It is useful to go deeper in this paper, because from this we took inspiration on how we could overcome some challenges we had to face during the design process of the system.

They have thought that IA should follow some rules in order to generate an acceptable and realistic environment. Some of these requirements includes: objects must have interaction zones or interaction standards, there must be a semantic coherence when the user interacts with them (he should be able to pick up an apple but not the water of a lake), a scene prompted must be coherent (if the user asked for a forest, the scene should include elements that are usually findable inside a forest). This is an important and difficult list of requirements.

They created a VR game with non-deterministic game mechanics powered by OpenAI’s text generative models, by integrating them with Unity game engine.

It is a sort of Ping Pong game where the players can transform both the paddles and the ball into any 3d objects and these transformed objects interact in semantically sensible; for instance, a ball transformed into an egg colliding with a pan result in a fried egg.

Going into the implementation of Codex VR Pong, the developers used an integration of Codex, which is a descendant of GPT-3 and its training data contains both natural language and billions of lines of source code from publicly available sources, on the other hand we have chosen the integration of ChatGPT in Unity for the output code.

The most important part of this scientific document is the way they used for overcoming the “problem” that C# code needs to be compiled, and Unity does not have a compiler that would allow evaluation of code at runtime. The solution to this problem is an open-source compiler called Roslyn downloadable from the Unity Asset Store; we will discuss in detail about this in the next chapter.

# Methods

In this chapter, we want to describe all the software and hardware used during the implementation. For each of them , we are going to give an idea of their working principles plus some technical aspects and specifications.

## Software

### Unity

### API OpenAI

### Roslyn C# Runtime Compiler

### Vocal Commands

## Hardware

### Meta Quest 2

# Design of the System

# Implementation

After the definition of the design of the system, we have to analyse and explain every single step of the implementation deeply. In this chapter, we provide a fully technical description of all the blocks needed to build the final framework.

# Results

## System Testing

### Experiments

### Script Generation: Execution Time

### Script Generation: Error Rate

## User Testing

### Testers Description

### System Usability Scale (SUS)

### User Experience Questionnaire (UEQ)

# Conclusions

## Future Works

Bibliography

[1] Morton Heilig. The cinema of the future. 1955.

[2] I. E. Sutherland, "Sketchpad-A Man-Machine Graphical Communication System", Proceedings of the Spring Joint Computer Conference, Detroit, Michigan, May 1963 (Washington, D.C.: Spartan, 1964).

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin “Attention Is All You Need” , June 2017

[4] <https://openai.com>, 2024

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei , “Language Models are Few-Shot Learners” , July 2020

[6] <https://chat.openai.com> , 2024

[7] Jasmine Roberts, Andrzej Banburski-Fahey, Jaron Lanier, “Steps towards prompt-based creation of virtual words”, November 2022