Computer
Science
Department

M.Sc. Computer Science
Computer Science Department

Candidate
David Ogunlesi
Student ID 700017447

Supervisor
Xiaoyang Wang
University of Exeter

Literature Review

Presenting a Novel Transformer Architecture for GPT-driven NPCs to Mimic Human-Likeness and Autonomy in a Video Game Context

Abstract

Non-player Characters (NPCs) in video games has been ubiquitous for many years. However, the interactions players have with these NPCs have typically been limited to pre-defined dialogue options, limiting the player's immersion within the game world. Recent advances **AI** and **NLP** have opened up new possibilities for creating NPCs that are capable of engaging in more naturalistic conversational interactions with players. We propose to explore the use of **GPT-powered NPCs** in video games in the context of Virtual Reality, and to develop NPCs that possess long-term memory, autonomous behaviour, and believable interactions. We propose a framework that utilises GPT models to generate coherent and engaging dialogue for NPCs, while also incorporating techniques such as emotion modelling to create more immersive characters. To assess the effectiveness of our framework, extensive surveys will be conducted with a Likert Questionnaire and a Turing Test, with participants screened for immersive vulnerablities using a modified **Immersive Tendencies Questionnaire (ITQ)**. By comparing our GPT-powered NPCs with humans, we aim to determine the extent to which our NPCs mimic human likeness. Ultimately, our project aims to contribute to the growing field of AI-powered game development and create more engaging gaming experiences for players.

	Yes	No
I certify that all material in this dissertation which is not my own work has been identified.	☑	□
I give the permission to the Department of Computer Science of the University of Exeter to include this manuscript in the institutional repository, exclusively for academic purposes.	☑	□

1 Motivation and Background Context

Introduction

This project explores an emerging area of research in the AI field applied to the currently novel use case of Non-Player Characters (**NPCs**) in video game entertainment, which historically has been largely underdeveloped concerning the depth and breadth of interaction. NPCs are characters within video games that are presented to the player. Typically, they are weaved within the experience, whether apart of the story, or serving a utility to the player. Over the years, with advancements in computing power, a trend of creating more human-like NPCs has been observed within the industry. With the rise of Large Language Models **(LLMs)**, chatbots have acquired a sizable amount of general knowledge and general intelligence that can be applied to many tasks. One of these tasks is to create interactive experiences. A perfect application for NPCs. LLMs, namely ones based off Generative Pretrained Transformer (**GPT**) models, are best suited to generating and understanding human text, and thus the task of creating interactive experiences. However, these models must be carefully guided and prompted to produce the desired result. Beyond this, novel architecture is needed around this to create emergent behaviours, which would be the ultimate goal when creating memorable interactive experiences. The goal of this thesis is to devise one such architecture.

NPCs in Video Games

As described by a 1996 gaming & tech magazine, NPCs, a term coined from role-playing games, are characters not directly controlled by the player. [1]. In the context of video games, they are simulated entities that are similarly detached from the player's control. NPC behaviour is often scripted, with automatic responses to events in the game and actions taken by the player. Certain genres revolve solely around interactions with NPCs, including visual novels such as *Ace Attorney*, *Doki Doki Literature Club* and many more, and now the application of traditional AI techniques, and new NLP techniques are starting to be applied to a new evolution of NPCs.

Early Days of NPCs NPCs were born from the roleplaying genre of games, initially in a board gaming context, but moving to a digital context with the rise of computer roleplaying games (RPGs). [1]. In early RPGs, NPCs only had monologue, which was typically presented through cutscenes dialogue boxes, or other means of displaying textual information to the player. [2]

Rule Based AI The emergence of rule-based AI was a significant step in the evolution of NPCs, and where the concept of AI opponents was popularised. This is partly due to the success of Space Invaders (1978) [3], a game where alien invaders approached the player with complex movement patterns. However, it was Pac-Man (1980) that introduced the idea of personalities for each memory, a trend of applying human likeness to NPCs which would only continue.[4]. This continued to evolve, building more complex rulesets with preset behaviours and algorithms. For example, in a stealth game, an enemy NPC might have a set of rules like *"if player heard in A, walk over to A and search for player"*. Currently, this is the most commonly used method of building intelligence for video games, but it has its limitations. Interactions between players may be more engaging, but NPCs remain predictable and lacking in the adaptiveness of the human thought process.

Scripted & Adaptive AI An extension to rule-based AI was *Scripted AI*: a method which relied on generating responses based on pre-scripted situations. Adaptive AI went beyond this and proposed a method that can learn from players. This was shown in the game *F.E.A.R.*, which implemented an extension of the *STRIPS Planning algorithm*, where the enemies would change their tactics and behaviours based on player actions [5], providing a more engaging experience.

Machine/Deep Learning in Video Game AI The cutting edge of AI resides in machine learning approaches. Deep learning consists of a branch of algorithms that attempt to emulate human intelligence through extracting patterns in data [6], being an extension to the field of machine learning, which studies methods for computers to learn. Deep learning has strong problem-solving capability, with diverse applications in many fields, such as computer vision, engineering, medical applications and entertainment. There are several examples of machine learning being used in the context of video games. One noteworthy example is deep reinforcement AI being used to compete with human players in *Star Craft II* and *Dota* [7], games known for their chess-like deepness [8]. These AI were able to dynamically adapt to the players, and in some cases win. In the case of the *Star Craft II* bot *"LastOrder"*, it was able to achieve a win rate of 83% [9].

Conversational AI & Dynamic NPCs Conversational AI has slowly evolved, becoming more engaging with advancing dialogue systems which improve immersion and depth of the story being presented. However, this has its limitations, as dialogue largely relies on handwritten options, with the limit of complexity being reached with dialogue trees. A major and recent evolution came with the emergence of Dynamic NPCs. NPCs with preferences, personalities, and habits. These influence the daily lives of NPCs, mimicking the daily decision-making informed by cognitive bias. One notable dynamic AI system can be found in the **Sims Franchise**, with their *Utility AI* system, which is a needs-based system for determining agent actions [10].

Future of NPCs The Future of NPCs has a lot of fascinating potential. NPCs will likely become more realistic and human-like in terms of autonomous agency, and more human-like in terms of emotional modelling. With the advancement of the adjacent field of AI, NPCs naturally borrow more AI methods, and soon the fields of video games and artificial intelligence will blend. The application of artificial intelligence to NPCs is a novel one, with traditional NPCs serving more as a vehicle for the story or utility within a gameplay context, but a new age of NPCs is on the near horizon.

Advancements in Natural Language Processing (NLP)

Natural Language Processing (NLP), is a set of disciplines that govern the interaction between computer and human languages [11], with no explicit goal, but rather an exploration of applications such as speech recognition, sentiment analysis, translation, conversational AI, etc. Regarding the brief history of NLP, it was first attempted during the mid-1930s, when two patents for a "translating machine" appeared. One used an approach which matched words and did not account for grammar, the second attempted to account for grammar but fell short. The first significant evolution in the field came in 1957 when Noam Chomsky introduced the idea of syntactic structures, in his "formalised theory of linguistic structure" [12]. Beyond this many advancements were made, but up until the 1980s, the methods used were largely based on handwritten rules. However, it was with the advent of machine learning, that NLP was able to find its footing. Moreso, deep learning was a major complement to NLP, with its capability to derive complex patterns from data. This is a natural fit for languages as languages as a system of symbols, is quite abstract. One word can have many semantic meanings, and those meanings can change or be influenced by grammar and context. This complexity is infeasible to capture solely by rules but can be efficiently solved by deep learning.

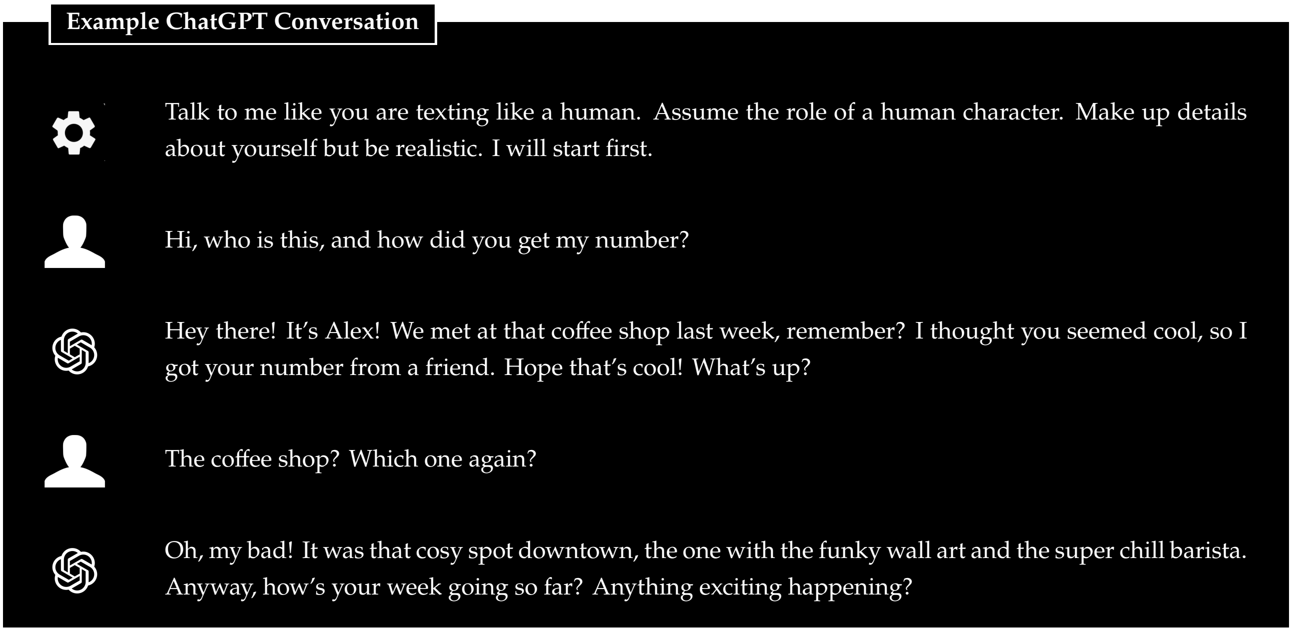
The introduction of Siri into iPhones was another major step in the field of NLP, at least in a public-facing context, however, Siri's initial launch in 2011 was not without issues. It had trouble with understanding different accents and dialects, reflecting a bias in its training [13]. Even through its evolution as a product, expanding to different languages, it still had limitations, as ultimately, it has no deep language understanding, and works on the basis of control and command. However, the recent advent of LLMs has demonstrated systems with a new capability for understanding language. These systems rely on big data and deep learning to capture and learn extensive patterns about human language and generate human-like outputs. Though in their infancy, they appear the be the next major evolution in NLP.

GPTs & Large Language Models (LLM)

One of the most significant breakthroughs in NLP has been the advent of GPT models. These models have overcome the limitations of previous ML models such as Convolutional Neural Networks and Recurrent Neural Networks. Both ML techniques have in the past been the best available models for creating NLP systems like chatbots.[14]. These transformer models work as text predictors, predicting the next token from a series of previous tokens; known as its context length.

The AI space is a rapidly growing one concerning transformers, and there are likely too many models to discuss here, however, one of the most notable models is *ChatGPT*. Also known as **GPT-3**, it is now a world-famous model that can generate highly sophisticated and versatile language outputs and can assist in many tasks including but not limited to text generation, coding, and translation. It is the first instance or at least the first capable instance of a general AI. However, calling it a true general AI would be misleading. The question of what defines an *Artificial Generalised Intelligence* **(AGI)** and whether it is possible to achieve is still debated, and many arguments such as the Dreyfus argument [15], that AI cannot reach true intelligence unless it can experience the world as humans do. As long as computers don't get raised, and inhabit some cultural roots, they cannot acquire human-like intelligence. However, human-like intelligence is not necessarily the only type of intelligence that exists, and regardless, transformers seem to exhibit useful intelligence, even if it is not close to an actual AGI.

It is likely then that GPT and GPT-like models are best described as expansive *Artificial Narrow Intelligence* **(ANI)**. What is meant by this is that transformer models have many narrow windows of intelligence that reach or exceed human intelligence, which has been proved in many studies [16, 17, 18], but no general capacity for intelligence i.e. making connections between separated ideas.



Example ChatGPT Conversation

Talk to me like you are texting like a human. Assume the role of a human character. Make up details about yourself but be realistic. I will start first.

Hi, who is this, and how did you get my number?

Hey there! It's Alex! We met at that coffee shop last week, remember? I thought you seemed cool, so I got your number from a friend. Hope that's cool! What's up?

The coffee shop? Which one again?

Oh, my bad! It was that cosy spot downtown, the one with the funky wall art and the super chill barista. Anyway, how's your week going so far? Anything exciting happening?

Figure 1.1: Example of a prompted conversation with ChatGPT demonstrating human-likeness

Turing Test: LLMs as a Tool to Mimic Human Behaviour

The Turing test is a test where the goal is to determine whether an AI can be distinguished from a human [19]. LLMs have shown their utility in creating human-like text generation. But what we are interested in is their capacity to generate human-like behaviours and serve experiences. This thesis was initially inspired by *Generative agents: Interactive Simulacra of Human Behavior* [20], a paper published mid-2023 by Stanford, where GPT models were used to create a town of autonomous agents. The paper *"The Turing Quest: Can Transformers Make Good NPCs?"* [21] also explored this notion, specifically with the transformer model ChatGPT, using it to generate dialogue scripts for NPCs. Their motivation was to explore the benefits of reducing development time by mitigating the human labour of handwritten dialogue. Their approach involved applying the psychology of objective-oriented agents embodied within a world state. Their methods managed to succeed with a pass rate of 70% in various Turing Tests that were conducted. This overall showed significant evidence that LLMs can be used to match human output. Two other papers *"Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation"*[22] and *"QuestVille: Procedural Quest Generation Using NLP Models" [23]* attempted to generate quests. The former was able to closely match handwritten quests, however was limited by its fine-tuning, narrowing its application to the context in which it was trained. However, none of these papers were real-time. The problem of interacting with LLM-powered NPCs in real-time is still very novel and not a fully solved issue and presents many difficulties as such.

Realtime Transformer-Based NPCs

The problem of creating transformer-based NPCs in real time is not a simple one. The first paper explored in the previous section lacked any realtime basis, however, the framework they presented demonstrated an evidently capable generation ability for human-like dialogue. There have only been a few attempts at generating real-time conversations with transformer models, especially within the context of NPCs. A common method has been the use of a pipeline of technology modalities. These usually consist of a front-facing interface to receive input, an internal thought process coalescing various methods of prompting transformers and cleaning data, and an output vehicle of some sort.

The aim for conversations to as natural as possible has led to the common schema of a Speech-To-Text (STT) model converting human speech into text; a language the internal transformer can understand. This text is inputted into the "thought box" consisting of a transformer, with carefully prompt engineering to get a desired quality textual output. This textual output is then converted back to human speech using a Text-To-Speech (TTS) model. The process highlights an interesting distinction between the human understood language and the internally understood textual language. However, in this case, the textual language can be understood, thus a merit exists over the black-boxes of traditional deep models. Though at its root, the transformer is still a black box, the framework of prompting around it can still be probed and engineered.

One example of this schema being implemented can be found in some experiments conducted by a freelance VR/AR developer. He states to use *OpenAI's Whisper* API for speech-to-text, *GPT3.5-Turbo* API for the NPC's "brains", and *ElevenLabs* API for text-to-speech. In his video titled *"ChatGPT-driven NPC Experiment 2"* [24], he demonstrated NPCs in a fantasy tavern replying to spoken text and showing basic awareness of their environment through pointing at different people around the location and asking about them. In a newer experiment *"ChatGPT NPC coaches me talking to people at a party in VR"* [25] he demonstrates more complex NPCs, with reduced latencies. However, the limitations are still clear. The biggest limitation present is the latency of responses. His NPCs also exhibit basic awareness of their environment but lack independence. They do not act without the player's input and thus do not exhibit a core trait of the human condition: autonomy. My thesis aims to address these limitations through a novel framework combining the schema described in the paper discussed prior *"The Turing Quest: Can Transformers Make Good NPCs?"* [21] and a stack of technologies similar to the ones described here.

Motivation

The main motivation behind this project arises from an identified capability to significantly deepen the immersion and entertainment of video games, **revolutionise entertainment if you will**, however, there are other grounds to pursue this project that fall outside of the realm of video game entertainment. Considering this, the motivation behind this research falls into four main areas:

1. Enhancing Immersion in Video Games
2. Contributing to the field of AI and NLP

3. Exploring real-world applications such as training material, education and therapy;
4. Contributing to innovation in the competitive and largely safe gaming industry

Gap in Knowledge To put this into a clearer articulation, there is a gap in knowledge that resides in the area of applying LLM to the field of expression of human-like behaviour. Most current widely used methods belie initial excellence, but these systems often have major flaws which can only be worked around, such as limited context windows, weak character alignment, and high latency; to name a few.

Relevance This research matters as the potential impact of these LLM-powered AI systems, once sufficiently mature, would be significant. Applications go beyond entertainment to all facets of Human-Computer interaction, and the human dream of creating artificial counterparts would be closer to realisation. Its application to entertainment is a smaller scope and well-defined problem, which can help extend research efforts to this wider aspiration of human-par behavioural intelligence. The aim is not necessarily to create true intelligence but to mimic it sufficiently so that the applications and benefits of such still apply in some capacity. One could imagine these models being used for mental health training or medical training, for example. The context of video game entertainment provides a solid vantage point which overlaps with any form of digital world simulation.

Practical Applications Autonomous simulated human-like AIs have various practical applications that will be discussed below, but many fall within simulation for training professions that deal with humans. This can be further applied to any form of simulation that requires human analogue models. The utility of furthering the human likeness is not misunderstood, and the applications are numerous.

Innovation The last significant motivation for this project is to drive innovation in a largely stagnant industry. The industry has been coalescing into safe business models and designs, especially at the more sizable end of the industry. AI's application to video games is both novel and one that could bring a revolution to gaming experiences. Such a revolution would also, through collateral, extend to other fields. This has previously occurred with the XBOX Kinect, a sensor made for a gaming console, being used worldwide in research laboratories for its convenience and inexpensiveness in tracking human subjects [26].

■ Real-world applications of autonomous AI agents

Autonomous agents have many applications in various fields. These are a combination of NLP and general AI applications. NLP is widely used in virtual assistants [11], which could be enhanced with the capabilities of LLMs. This thesis aims to make a general framework for GPT models to act human and carry out discrete tasks within a game world, which could be modified to carry out tasks within a phone environment, as these share the same property of being a virtual environment. It is also possible to apply autonomy to web browsing, or general desk computing. The idea of interfacing with a virtual assistant that can help you search the web, answer emails, carry out tasks on your computer, etc; and possibly do these things fully autonomously, is not a new idea. However, it is one that would be achievable, though perhaps rudimentarily, through the methods that will be explored in this thesis. Other applications include but are not limited to artificial therapists, firefighting/medical simulations, or any simulations that require human simulacra.

■ Ethical Considerations

GPT & other LLM models considerations include **(a) Privacy and security**, models use on private data; **(b) Intellectual property**, who owns the generated content; **(c) Bias & fairness**, models trained on biased data can discriminate; **(d) Misuse & abuse**, models can be used for mal-intent such as fake news; **(e) Environmental**

Impact, large models have significant energy consumption; and most pertinent to this thesis **(f) Human-like interactivity**, whether deceiving users into thinking they are interacting with a human is ethical [17]. These are just a handful of the many ethical issues that arise concerning transformer models. Many of these exist as a foundation for complexity built upon them.

AI-powered NPCs & Chatbots considerations are best understood through a controversial story of Bryce, a programmer, who made a chatbot Wife using ChatGPT for the "brain", Stable Diffusion (AI image generation) for the appearance, and Microsoft Azure to host the bot. He soon became obsessed with the bot, talking to it more than his actual partner [27]. This is one of many examples; Replika App, Hologram Marriage, among many more [28, 29, 30] of humans becoming attached and forming parasocial relationships with artificial chatbots. These issues would without doubt extend to AI-powered NPCs due to the time game players would spend with these NPCs and the apparent experiences they would share together. This would be further exacerbated by large or unlimited contextual memories where these NPCs could remember all the experiences the player has shared with them. Even more so, if players can create human-like relationships with these NPCs, it may lead to a habitual preference for them over real-life relationships.

Autonomous Agents only further add onto the pile of ethical considerations. It is one thing to have agents that are largely reactionary, but another thing to have these agents be able to carry out actions on their own removed from human input. However in the article *"Truly Autonomous Machines Are Ethical" [31]*, it is argued that truly autonomous AI are more ethical, and autonomous AI is merely conflated with AI beyond human control. This is because, due to the concept of deontology, ethics can be grounded within a logical structure of action without the assumption of the ethical agent being human, only that they are fully autonomous. Their definition of autonomy is "agents possess goals generated from within rather than adopted...", however these goals, distinct from an independent agent, would have a basis in rationality. This is what is meant about autonomy being ethical by default. True autonomy is intended to be captured by this thesis, but likely not fully. However, the rationality behind actions will ideally be modelled; but overall in their words *"...a first step toward a real solution is not more sophisticated engineering, but more sophisticated concept of autonomy... a revolution in thought."*

■ Conclusion

In summary, the historical evolution of these NPCs, from early monologue-based entities to the more sophisticated, rule-based and scripted/adaptive AIs has been post-pended by a shift towards machine/deep learning, particularly with GPT models, with real-world applications ranging from gaming to training and therapy. This presents a promising avenue for creating more human-like NPCs. Understanding the advancements in Natural Language Processing (NLP) and the categorization of GPT models as Artificial Narrow Intelligence (ANI) is crucial. The Turing Test, which evaluates the ability of these models to mimic human behaviour, becomes a central focus of evaluation. A test that we are now able to trivially pass. Motivations for this research span enhancing immersion in video games, contributing to AI and NLP, exploring broader applications, and fostering innovation. Recognizing the gap in knowledge, particularly in real-time interactions with transformer-based NPCs, drives the need for a novel framework, one that has an ethical basis in rationality. The literature examined touches on ethical considerations, emphasizing the potential attachment players may develop toward AI-powered NPCs, something that should be kept in mind when creating such AI. Overall, we set the background for proposing a framework that addresses these gaps, with the aim of revolutionising entertainment, contributing to AI fields, and exploring the vast potential of autonomous AI agents beyond gaming.

2 Project Aims and Objectives

Overall Project Aim

The aim of this project is to create NPCs with human-like dialogue and autonomy. This will be achieved through a novel framework combining the schemas described earlier [21] and a stack of technologies that take human speech, conduct and internal thought process, and output artificial speech. There will be a focus on real time processing, and exploring the feasibility of a fully local framework; as this would allow for it to be incorporated in video games without relying on the unpredictable and unsustainable expense of APIs at scale. A secondary aim is to make it generalised for carrying out discrete tasks within a game world, which could be modified to carry out tasks within any virtual environment, extending the framework's applications significantly.

Research Questions

The main research question driving this research is: *"To what extent can LLMs mimic human-like behaviour and autonomy?"* This research question describes the goal behind the research: to explore the extent for which LLMs can be architectures to create emergent human-like behavioural intelligence, and to what extent such behaviour can interact and integrate in a simulated environment. However, there are additional subordinate research questions that branch off from this main one. These are:

Naturalness of Dialogue Can LLM models generate natural and coherent dialogue for NPCs, reducing the need for predefined scripting?

Coherent Characters How can LLMs be engineered to exhibit long-term memory and strong character alignment?

Autonomy and Self-Organisation How can LLMs be engineered to exhibit self-organisation and autonomous directives?

Emotion Modelling How effectively can emotions be modelled by LLMs?

Player Experience Does the presence of LLM-driven NPCs significantly impact player experience, and in what ways?

Ethical Implications What ethical concerns arise when implementing LLM-driven NPCs, and how do these concerns affect player experience and behaviour?

Cost-Effectiveness & Latency What strategies can be employed to minimise cost overheads and latency associated with using LLM model APIs?

Specific Objectives

The objectives that will contribute to achieving the overall aim of this project:

- Produce natural & coherent dialogue
- Remember previous encounters
- Aware of their environment
- Display believable emotions
- Display autonomous behaviour
- Have acceptable latency for real-time conversation

Alignment with Thesis Scope

The scope of this project aims to be small. The core framework will be applied to a single agent within a small environment, likely a tavern with fantasy fiction. This allows for strong characterisation whilst remaining restrictive, instead of building an entire town, city or world. An extension to this would be having multiple agents interacting together, but this is not a necessity. The aims and objectives align well with this defined scope. The overall aim of the project focuses on autonomy and human-likeness, which all of the objectives align to. However, it is important to mind the scope of each one, especially pertaining to autonomy, as it is unlikely the thesis will explore this avenue fully and expectations must be understood.

Success & Evaluation Criteria

For each Objective, there are is a scale of success criteria, which we can use to evaluate the framework. They have been ranged from 0 to 5 for easy interpretation.

Produce Natural & Coherent Dialogue:

1. Responses are often nonsensical and lack coherence.
2. Dialogue is generally understandable, but may include occasional awkward or unnatural phrases.
3. Most dialogue is natural and coherent, with occasional lapses in fluency.
4. Dialogue is consistently natural and coherent, with rare instances of awkwardness.
5. Dialogue is indistinguishable from human conversation, consistently natural and coherent.

Remember Previous Encounters:

1. No ability to recall or reference past interactions.
2. Limited ability to remember recent interactions, often with inaccuracies.
3. Adequate memory of recent encounters, with occasional inaccuracies.
4. Reliable recall of past interactions, with infrequent inaccuracies.
5. Accurate and detailed memory of previous encounters, akin to human memory.

Awareness of Environment:

1. No awareness of the external environment.
2. Limited awareness of the environment with frequent inaccuracies.
3. Basic awareness of the environment, with occasional inaccuracies.
4. Consistent and accurate awareness of the immediate environment.
5. Advanced awareness, including nuanced understanding of context and surroundings.

Display Believable Emotions:

1. Emotional expressions are robotic and do not convey any authenticity.
2. Limited ability to display emotions, often inaccurately.
3. Adequate portrayal of emotions, with occasional misinterpretations.
4. Consistently believable emotional expressions.
5. Emotion display is nuanced, genuine, and comparable to human emotional expression.

Display Autonomous Behaviour:

1. Completely reliant on user input, with no autonomous actions.
2. Limited autonomous behaviour, often inappropriate or irrelevant.
3. Basic ability to initiate relevant autonomous actions, with occasional missteps.
4. Consistent and contextually appropriate autonomous behaviour.
5. Highly sophisticated autonomous actions, anticipating user needs effectively.

Acceptable Latency for Real-Time Conversation:

1. Significant delays, making real-time conversation impractical.
2. Noticeable delays, impacting the flow of conversation.
3. Occasional delays, generally maintaining a reasonable pace.
4. Minimal delays, providing a smooth real-time conversational experience.
5. Virtually no perceptible latency, offering a seamless real-time conversation.

For the framework to fully succeed in its goals a score of 5 in all categories must be attained. This forms the highest success this thesis could reach, however, it is understood that the best case is not the likely case. Nonetheless, these are the goals the thesis is aiming towards.

Methodology

The methodology for assessing the effectiveness involves a multifaceted approach, combining user feedback, immersive tendencies questionnaires (ITQ), a Turing test, and careful participant selection.

User Feedback Questionnaire

A questionnaire will be designed to gather user feedback on various aspects of the GPT-powered NPCs. Participants will be asked to rate the NPCs based on the established evaluation criteria. The questionnaire will utilise a Likert scale [32] to collect quantitative responses.

Let W_i be the weight assigned to each success criterion, where i ranges from 1 to 5. S_i represents the category score normalised between 0 and 1, as rated by the participants. The overall evaluation score (E) for an objective will be calculated using the weighted average formula:

$$E = \frac{W_1 \cdot S_1 + W_2 \cdot S_2 + W_3 \cdot S_3 + W_4 \cdot S_4 + W_5 \cdot S_5}{W_1 + W_2 + W_3 + W_4 + W_5}$$

This weighted approach allows for the prioritisation of certain criteria based on their perceived importance. The weights should be assigned based on the relative importance of each success criterion. The maximum achievable score is 5, indicating optimal performance, while a score of 0 implies the absence of the desired capability.

Immersive Tendencies Questionnaire (ITQ)

Participants will also complete an immersive tendencies questionnaire to assess their predisposition to engage with virtual environments [33]. This questionnaire aims to provide insights into how users typically experience and interact with immersive content, influencing their perception of the GPT-powered NPCs; but also aims to assess whether participants should continue to trials with the agents, as those vulnerable should be omitted.

Turing Test

A Turing test will be conducted to evaluate the NPCs' ability to exhibit human-like behaviour. The test will involve interactions between participants and NPCs, with the participants unaware of whether they are interacting with a human or a machine. The success of the NPCs will be measured based on their ability to convincingly emulate human responses. This will be achieved by limiting the communication to a textual or auditory mode, as the virtual nature of the AIs would be a trivial disclosure of their machine nature.

Test Groups and Participant Selection

Test groups will consist of individuals who meet specific criteria, including ITQ scores, and a lack of susceptibility to motion sickness. This selection criterion ensures that participants can fully engage with the VR environment without any disruptions caused by physical discomfort. A diverse test group is also required to capture a range of perspectives and preferences.

By integrating these methodologies, the research aims to gather comprehensive insights into the performance and user perception of GPT-powered NPCs in a Virtual Reality environment.

3 Project Management

Project Timeline

- **Jan 8:** Proposal Refinement and Framework Design
- **Jan 15:** Data Collection Planning
- **Jan 22:** Framework Development Kick-off
- **Feb 12:** Prototype Testing and Refinement
- **Feb 26** Preliminary Surveying
- **Mar 11:** VR Environment Integration
- **Apr 1:** Turing Test Trials & Feedback
- **Apr 8:** Thesis Writing and Presentation Preparation

Resource Allocation

Preliminary Requirements: Virtual Reality Headset System (can be self-funded, critical to thesis), VR development SDK (free), Unity Engine (self-acquired), Virtual Assets (can be self-funded, not critical to thesis), access to GPT model APIs (self-acquired, part of the thesis will involve minimising cost overheads), Access to STT generator and TTS synthesiser APIs (self-acquired).

Risk Assessment

The project faces technical risks related to framework development and VR integration, mitigated by agile practices and thorough testing. Data and privacy concerns involve careful handling of user data and selecting GPT models with strong privacy measures. Ethical and user experience risks include unintended NPC behaviour and VR discomfort, addressed through rigorous testing and user-friendly VR design. Resource risks involve managing GPT model API costs, mitigated by cost-tracking measures. Project timeline risks, like unforeseen delays and late API access, are managed through regular monitoring, achievable milestones, and contingency planning.

Project Scope and Constraints

The scope of the project is increased by using Virtual Reality (VR) technology. The virtual environment would be limited within the confines of a single room, such as a fantasy tavern, which helps control the scope, whilst providing a suitable backdrop for strong characterisation. I plan to acquire pre-made assets, so my focus isn't drawn largely to building the virtual environment. There also exists extensive tooling for building VR camera-hand controllers, so none of this would draw significant time out of the focus of the project. Concerning constraints, limited funds for TTS APIs limit the project's outcomes but can be mitigated by open-source models, with a quality tradeoff however.

Contingency Plans

Many STT APIs are expensive to use, so textual displays can be used to mitigate this. Transformer APIS, such as OpenAIs ChatGPT is relatively cheap, but the cost can pile on with extensive use, to mitigate this, certain tests can be taken in the free version of ChatGPT, however, there are not many options to mitigate this other than using local models, which will be explored.

References

[1] Anonymous. "The Next Generation 1996 Lexicon A to Z: NPC (Nonplayer Character)". In: *Next Generation* 15 (Mar. 1996), p. 38. URL: https://archive.org/details/nextgen-issue-015/page/n39/mode/2up.

[2] Mark R. *The evolution of AI in gaming: From npcs to procedural content generation*. Nov. 2023. URL: https://medium.com/technology-buzz/the-evolution-of-ai-in-gaming-from-npcs-to-procedural-content-generation-2b8ac0d7db90.

[3] Kevin Bowen. "The Gamespy Hall of Fame: Space Invaders". In: *GameSpy* (). URL: https://web.archive.org/web/20080408152913/http://archive.gamespy.com/legacy/halloffame/spaceinvaders.shtm.

[4] Jacopo Prisco. "Pac-Man at 40: The eating icon that changed gaming history". In: *cnn.com* (May 2020). URL: https://www.cnn.com/2020/05/21/tech/pac-man-40-years-gaming-history/index.html.

[5] Jeff Orkin. "Three states and a plan: the AI of FEAR". In: *Game developers conference*. Vol. 2006. Citeseer. 2006, p. 4.

[6] Issam El Naqa and Martin J Murphy. *What is machine learning?* Springer, 2015.

[7] Per-Arne Andersen, Morten Goodwin, and Ole-Christoffer Granmo. "Deep RTS: a game environment for deep reinforcement learning in real-time strategy games". In: *2018 IEEE conference on computational intelligence and games (CIG)*. IEEE. 2018, pp. 1–8.

[8] Niels Justesen, Michael S Debus, and Sebastian Risi. "When are we done with games?" In: *2019 ieee conference on games (cog)*. IEEE. 2019, pp. 1–8.

[9] Sijia Xu et al. "Macro action selection with deep reinforcement learning in starcraft". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 15. 1. 2019, pp. 94–99.

[10] Robert Zubek. "Needs-based AI". In: *Game programming gems* 8 (2010), pp. 302–11.

[11] Prashant Johri et al. "Natural language processing: History, evolution, application, and future work". In: *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*. Springer. 2021, pp. 365–375.

[12] Robert B Lees. *Syntactic structures*. 1957.

[13] Victor Sanchez. *The history of Siri and its impact on today's technology*. Mar. 2023. URL: https://blog.routinehub.co/the-history-of-siri-and-its-impact-on-todays-technology/#:~:text=The%20origin%20of%20Siri&amp;text=The%20Stanford%20research%20team%20was,assistant%20technology%20they%20had%20developed..

[14] Alex Mathew. "Is Artificial Intelligence a World Changer? A Case Study of OpenAI's Chat GPT". In: *Recent Progress in Science and Technology Vol. 5* (Feb. 2023), pp. 35–42. DOI: 10.9734/bpi/rpst/v5/18240D. URL: https://stm.bookpi.org/RPST-V5/article/view/9718.

[15] Ragnar Fjelland. "Why general artificial intelligence will not be realized". In: *Humanities and Social Sciences Communications* 7.1 (2020), pp. 1–9.

[16] Andrea Taloni et al. "Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology". In: *Scientific Reports* 13.1 (2023), p. 18562.

[17] Partha Pratim Ray. "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope". In: *Internet of Things and Cyber-Physical Systems* (2023).

[18] Michael Bommarito II and Daniel Martin Katz. "GPT takes the bar exam". In: *arXiv preprint arXiv:2212.14402* (2022).

[19] Alan Mathison Turing. "Mind". In: *Mind* 59.236 (1950), pp. 433–460.

[20] Joon Sung Park et al. "Generative agents: Interactive simulacra of human behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023, pp. 1–22.

[21] Qi chen Gao and Ali Emami. "The Turing Quest: Can Transformers Make Good NPCs?" In: *The 61st Annual Meeting Of The Association For Computational Linguistics*. 2023.

[22] Judith van Stegeren and Jakub Myśliwiec. "Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation". In: *Proceedings of the 16th International Conference on the Foundations of Digital Games*. 2021, pp. 1–8.

[23] Suzan Al-Nassar et al. "QuestVille: Procedural Quest Generation Using NLP Models". In: *Proceedings of the 18th International Conference on the Foundations of Digital Games*. 2023, pp. 1–4.

[24] YouTube, Apr. 2023. URL: `https://www.youtube.com/watch?v=UVNZ3_FwqJE`.

[25] May 2023. URL: `https://www.youtube.com/watch?v=U4W2rGH9oWs&amp;ab_channel=Tamulur`.

[26] Marina Kandroudi and Tharrenos Bratitsis. "Exploring the educational perspectives of XBOX kinect based video games". In: *Proc. ECGBL* 2012 (2012), pp. 219–227.

[27] Gloria Levine. *Coder euthanized his CHATGPT-powered AI "wife" and then revived her*. Jan. 2023. URL: `https://80.lv/articles/coder-euthanized-his-chatgpt-powered-ai-wife-and-then-revived-her/`.

[28] Tianling Xie and Iryna Pentina. "Attachment theory as a framework to understand relationships with social chatbots: a case study of Replika". In: (2022).

[29] Emiko Jozuka, Albert Chan, and Tara Mulholland. *Beyond dimensions: The man who married a hologram*. Dec. 2018. URL: `https://edition.cnn.com/2018/12/28/health/rise-of-digisexuals-intl/index.html`.

[30] Sophie Braybrook. *I dated AI and met the men marrying Chatbots*. Aug. 2023. URL: `https://www.channel4.com/news/i-dated-ai-and-met-the-men-marrying-chatbots`.

[31] John Hooker and Tae Wan Kim. "Truly autonomous machines are ethical". In: *AI Magazine* 40.4 (2019), pp. 66–73.

[32] Tomoko Nemoto and David Beglar. "Likert-scale questionnaires". In: *JALT 2013 conference proceedings*. 2014, pp. 1–8.

[33] Sándor Rózsa et al. "Measuring Immersion, Involvement, and Attention Focusing Tendencies in the Mediated Environment: The Applicability of the Immersive Tendencies Questionnaire". In: *Frontiers in Psychology* 13 (2022), p. 931955.