T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF COMMUNICATION DESIGN

EMPOWERING NARRATIVE DESIGN IN VIDEO GAMES WITH LARGE LANGUAGE MODELS: A CASE STUDY OF ASSASSIN'S CREED ODYSSEY

MASTER'S THESIS

TAMER UMUT KELTEK

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF COMMUNICATION DESIGN

EMPOWERING NARRATIVE DESIGN IN VIDEO GAMES WITH LARGE LANGUAGE MODELS: A CASE STUDY OF ASSASSIN'S CREED ODYSSEY

MASTER'S THESIS

TAMER UMUT KELTEK

THESIS ADVISOR

PROF. DR. BARBAROS BOSTAN

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	Game Design Master's Program
Student's Name and Surname:	Tamer Umut Keltek
Name Of The Thesis:	Empowering Narrative Design in Video Games with Large Language Models: A Case Study of Assassin's Creed Odyssey
Thesis Defense Date:	03.09.2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Yücel Batu SALMAN

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Institution**	**Signature**
Thesis Advisor's	Prof. Dr. Barbaros Bostan	Bahçeşehir University	
Member's	Dr. Öğr. Üyesi Ertuğrul Süngü	Bahçeşehir University	
Member's	Doç. Dr. Uğur Kaplancalı	Yeditepe University	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: **Tamer Umut Keltek**

Signature:

ABSTRACT

EMPOWERING NARRATIVE DESIGN IN VIDEO GAMES WITH LARGE LANGUAGE MODELS: A CASE STUDY OF ASSASSIN'S CREED ODYSSEY

Tamer Umut Keltek

Master's Program in Game Design

Supervisor: Prof. Dr. Barbaros Bostan

August 2024, 68 pages

This thesis explores the innovative application of Large Language Models (LLMs) in video game quest design, with a specific focus on Assassin's Creed Odyssey. The study aims to develop a methodology for fine-tuning LLMs using synthetic data derived from structured quest information and to evaluate their effectiveness in generating contextually appropriate quests compared to non-fine-tuned models. Leveraging community-driven content from the Assassin's Creed Fandom Wiki, the research constructs a robust dataset to fine-tune models like ChatGPT 3.5 Turbo and GPT-4o-mini. The findings reveal that while fine-tuning offers potential for specialized content generation, base models often outperform fine-tuned variants across multiple evaluation criteria. The research highlights the importance of structured data and nuanced fine-tuning strategies in achieving high-quality, AI-generated content. This work contributes to the growing body of knowledge on the intersection of AI and game design, offering insights into the future of narrative generation and the role of AI in creative industries.

Keywords: Large Language Models, Artificial Intelligence, Procedural Content Generation, Game Design, Game Evaluation Criteria

ÖZ

BÜYÜK DİL MODELLERİ İLE VİDEO OYUNLARINDA HİKAYE TASARIMINI GÜÇLENDİRMEK: ASSASSIN'S CREED ODYSSEY ÜZERİNE BİR VAKA ÇALIŞMASI

Tamer Umut Keltek
Oyun Tasarımı Yüksek Lisans Programı
Tez Danışmanı: Prof. Dr. Barbaros Bostan

Ağustos 2024, 68 sayfa

Bu tez, video oyunlarında bulunan görevlerin, büyük dil modelleri aracılıyığıyla tasarlanması ile ilgili olan yenilikçi uygulamaları Assassin's Creed Odyssey örneği üzerinden incelemektedir. Bu çalışma, topluluk tarafından gönüllü bir şekilde kayıt altına alınmış oyun verilerinin, sentetik bir veritabanı haline getirilerek büyük dil modellerinin fine-tune edilmesi üzerine bir metodoloji kurgulamaktadır. Farklı parametreler ile fine-tune edilmiş modellerin ürettiği görevler farklı kriterler doğrultusunda değerlendirilmiş ve bu modellerin bağlamsal olarak uygun görevler üretme konusundaki yetkinlikleri karşılaştırılmıştır. Araştırmanın sonuçları, fine-tuned edilmiş modellerin içerik üretimi konusunda potansiyel sunduğu ancak temel modellerin bir çok değerlendirme kriterinde fine-tuned edilmiş varyantlara göre daha iyi performans gösterdiğini ortaya koymaktadır.

Anahtar Kelimeler: Büyük Dil Modelleri, Yapay Zeka, Prosedürel İçerik Üretimi, Oyun Tasarımı, Oyun Değerlendirme Ölçeği

ACKNOWLEDGEMENTS

I extend my deepest gratitude to my mother, Yasemin Keltek, and my father, Mustafa Keltek, for their boundless love and support. Their unwavering support has been a source of strength for me.

I would also like to dedicate a special acknowledgement to my late grandmother, whose memory and spirit have been a guiding light for me. Her legacy of resilience continues to influence me deeply and this achievement is a tribute to her loving memory.

I am immensely grateful to Prof. Barbaros Bostan for his invaluable guidance and mentorship throughout this research.

This journey would not have been possible without each of you, and I am deeply appreciative of the support provided.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
MVP	Minimum Viable Product
NLP	Natural Language Processing
GaaS	Games as a Service
PCG	Procedural Content Generation
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
DBN	Deep Belief Network
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
QLoRA	Quantized Low-Rank Adaptation
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
F1 Score	Harmonic Mean of Precision and Recall
GPT-3	Generative Pre-trained Transformer 3
GPT-4	Generative Pre-trained Transformer 4
XML	eXtensible Markup Language
JSON	JavaScript Object Notation
CC-BY-SA	Creative Commons Attribution-ShareAlike
USMLE	United States Medical Licensing Exam

Chapter 1

Introduction

Humanity is on the verge of a major transformation, primarily fueled by progress in Large Language Models (LLMs), a branch of artificial intelligence (AI). These technologies signal a new era where interdisciplinary methodologies and applications become seamlessly integrated, enhancing research, analysis, and discovery. While some critics argue that this integration could compromise research credibility due to a lack of deep methodological expertise, this conservative viewpoint overlooks the innovative potential of these advancements.

From the earliest ancestors of humankind to our modern society, storytelling has been a vital means of preserving and transmitting knowledge to ensure our future generations endure through time. Stories have been used to explain life, preserve history, and ensure the continuity of experiences (Abrahamson, 1998). Initially, they were about transferring information that would increase the rate of survival. Then, especially starting with the Age of Enlightenment, those stories became more sophisticated, shaping how humans live and perceive life.

Most of the stories are conveyed through language, a distinctive human capacity for expression and communication that develops early in life and evolves over a lifetime (Pinker, 2014; Hauser et al., 2002). Similar to the evolution of language for humans, stories emerged with the development of culture and technology. As societies grew and advanced, so did the complexity and richness of their narratives. The progression from oral to written to printed and ultimately digital methods of communication represents a fundamental shift, impacting not only how stories are told but also how knowledge is processed and understood which signifies a deeper shift in human consciousness (Ong, 1982).

That shift has become more apparent with the concept of "new media," which has changed the landscape of storytelling and knowledge dissemination significantly.

New media allows human beings to transmit different narratives between different receivers asynchronously, creating a chaotic network of information flow. Today, that chaotic network becomes even more complex with the advancements in Large Language Models (LLMs), as it allows people to create their own narratives anytime based on their imagination by communicating with systems (machines) that mimic human language, using a probabilistic approach in the background. This development represents a significant leap in the evolution of storytelling, further blurring the lines between author, audience, and medium.

As storytelling evolves through digital enhancements, traditional narrative structures become less competent. The interactive elements of digital mediums introduce a layer of uncertainty to static scaffolds, allowing for more dynamic and personalized storytelling experiences. This phenomenon becomes more noticeable in the context of video games because of their interactive nature which converts the audiences from passive observers to active participants who are able to change the "play-ground" within the limits the game allows (Huizinga, 1950).

The participatory nature of video games creates immersive subjective experiences for individuals and challenges traditional narrative structures, ultimately establishing them as a dominant cultural form (Muriel & Crawford, 2018). Indeed, video games have emerged as a major artistic and cultural force, both influencing and being influenced by other media such as cinema, literature, theatre, and music. This interconnectedness underscores the importance of studying games within the broader context of contemporary culture; ignoring their impact and evolution risks a narrow understanding of our current cultural dynamics (Muriel & Crawford, 2018).

In addition to their importance of understanding cultural dynamics, video games occupy a unique meeting point between social sciences and analytical disciplines. This convergence creates an opportunity to explore the interaction of diverse methodologies to derive valuable insights. Given that the parameters influencing video games originate from various fields, game design offers substantial room for exploration. By studying the multifaceted essence of video games, researchers can gain deeper insights

into human behavior, cultural trends, and technological advancements, solidifying game design as a vital area for academic and practical inquiry.

The evolution of storytelling and the rise of video games as a dominant cultural form have set the stage for new approaches to creating narratives and game designs. As games become increasingly complex and immersive, the demand for rich, engaging content grows exponentially. This is particularly evident in dynamic and open-ended games, where players' demand for more game content continues to rise (Merrick, 2008). Quest design, a crucial element that drives player engagement and narrative progression, exemplifies this growing need for compelling content.

Quest design presents unique challenges which demand a delicate balance of creativity, narrative coherence and gameplay mechanics. Designers must craft experiences that are both compelling and varied, while ensuring they fit seamlessly into the larger game world and overarching storyline. This process is often resource-intensive, leading to a growing need for innovative solutions that can streamline and enhance the quest creation process.

Large Language Models (LLMs) have shown great capabilities in understanding and generating human-like text. Their ability to process and produce complex narrative structures with the given input, present an intriguing possibility for augmenting the quest design process. By leveraging the vast amount of existing game data and the generative capabilities of LLMs, there is potential to create tools that could assist game designers in developing diverse, creative and engaging quests more efficiently.

However, the application of LLMs in game design is not without challenges. While general purpose language models are powerful, their lack of the specific knowledge and nuanced understanding required for effective quest design. This is where the concepts of fine-tuning and retrieval-augmented generation (RAG) comes to the equation. By training these models on specific, high-quality quest data from successful games, we can potentially create specialized tools that understand intricacies of quest design and can generate relevant, context-appropriate content.

This study explores the innovative application of Large Language Models (LLMs) in the context of video games quest design with a specific focus on Assassin's Creed Odyssey. It aims to develop a methodology for fine-tuning LLMs using synthetic data derived from structured quest information and evaluate their effectiveness in generating contextually appropriate quests compared to non-fine-tuned models. By using synthetic data created from fan-contributed content on the Assassin's Creed Fandom Wiki, this study investigates the potential of community-driven information transformed into structured, machine-learning-ready datasets for advancing AI applications in game development.

This study sits at the intersection of game design, narrative theory, and artificial intelligence, offering insights into how advanced language models might be leveraged in the creative processes such as game development. By exploring this confluence of technology and creativity, it is hoped to contribute to the ongoing dialogue about the role of artificial intelligence (AI) in creative industries and its potential to augment, rather than replace, human creativity in game design. This research extends beyond the narrative of games; it explores how methodologies from various disciplines can be applied accurately, sustainably, and effectively in the rapidly evolving landscape of large language models (LLMs) and game development.

Chapter 2

Literature Review

This chapter provides an overview of the evolution of artificial intelligence, focusing on the development of large language models (LLMs) and their application in narrative generation and video game design. It explores the historical foundations of AI, the integration of AI in storytelling, and the specific use of AI in enhancing quest design within games, with a particular emphasis on Assassin's Creed Odyssey. Additionally, it touches on the use of synthetic data in fine-tuning AI models and considers the future implications and ethical considerations of AI-assisted game development.

2.1 Evolution of Artificial Intelligence: From Early Models to LLM

2.1.1 From early AI to neural networks. Throughout history, humanity has drawn inspiration from the natural world, particularly from its own intricate biological processes. This inclination toward self-reflection and emulation is evident in the early stages of artificial intelligence (AI), where pioneers endeavoured to mimic human cognitive functions through mathematical and computational models. A significant example of this effort is the pioneering work of neurophysiologist Warren McCulloch and logician Walter Pitts, who, in 1943, developed the first mathematical model for neural networks. Their seminal work, titled "A Logical Calculus of the Ideas Immanent in Nervous Activity," was inspired by the human brain's functioning, especially how neurons process information. This model marked the first mathematical representation of a biological neuron, a concept that continues to resonate in computational neuroscience and artificial intelligence.

McCulloch and Pitts aimed to capture the essence of neural computation through a simplified, binary model of a single neuron (Pospíchal & Kvasnicka, 2015). Their model posited that neurons receive inputs from other neurons, process these inputs, and then transmit outputs to subsequent neurons. Mathematically, this was represented

by each artificial neuron receiving multiple binary inputs. The neuron would compute the sum of these inputs and, if the total exceeded a predetermined threshold, the neuron would "fire" and produce an output; otherwise, it would remain inactive (Anderson, 2006).This approach was grounded in the "all-or-none" law of nervous activity, which suggests that any neuron's activity can be expressed as a proposition (McCulloch & Pitts, 1943).

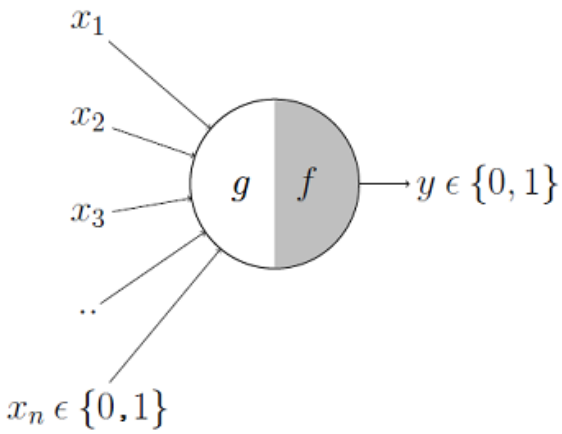

Figure 1 A simplified model illustrating the logical structure of early artificial neural networks.

Despite its simplicity by modern standards, the McCulloch-Pitts model played a crucial role in the development of more advanced artificial neural networks. It introduced the concept of fixed binary values for neuron activation, which anticipated the adjustable weights in contemporary networks that enhance learning through adaptation. This early work established threshold functions that are essential in today's neural architectures and demonstrated that networks could perform basic logical operations (AND, OR, NOT), thus laying the foundation for computational logic (Pospíchal & Kvasnicka, 2015). The significance of this model extends beyond its technical achievements; the interdisciplinary collaboration between McCulloch and Pitts was instrumental in advancing cybernetics and cognitive science (Abraham, 2002). This interdisciplinary approach set a precedent for integrating diverse fields to drive innovation in AI, a principle that continues to be a cornerstone of modern research. Today's deep learning models, including Large Language Models like GPT-3, can be viewed as sophisticated descendants of this early concept, employing billions

of interconnected neurons to process and generate human-like text. The progression from the McCulloch-Pitts model to contemporary AI systems, which leverage biases to shift activation functions and fine-tune decision-making processes, illustrates how the convergence of diverse disciplines can lead to robust, innovative solutions that advance human understanding.

As research continued into the 1950s, the focus transitioned from theoretical models to observable behavior. A pivotal moment in this shift was Alan Turing's introduction of the Turing Test in his seminal paper "Computing Machinery and Intelligence." This innovative evaluation method, often referred to as the "imitation game," involves a human judge interacting with both a machine and another human via a text-based interface, with the aim of identifying which is the machine. Turing argued that if the judge could not reliably distinguish between the two, the machine should be considered to exhibit intelligent behavior (Turing, 1950). This approach reframed the criterion for intelligence, shifting it from internal cognitive processes to observable outcomes, bypassing philosophical debates about consciousness. The Turing Test had profound implications, moving AI research towards replicating human-like behavior and providing a clear, practical criterion for evaluating machine intelligence. As Turing stated, "The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavor that we wish to include," emphasizing the test's broad applicability (Turing, 1950). This pragmatic approach allowed for empirical validation and catalyzed research efforts towards developing machines that could convincingly simulate human conversation and interaction.

Interestingly, Turing's era was concerned with machines exhibiting human-like intelligence. Today, the challenge has evolved: we now strive to distinguish human-generated content from AI-generated content. AI detectors often struggle to accurately identify whether a text was written by a human or a machine, underscoring the complexity and sophistication of modern AI systems. This contemporary issue highlights the progress and challenges in AI, as the boundary between human and machine-generated content becomes increasingly blurred.

The mid-20th century also witnessed key advancements that paved the way for the formal recognition of artificial intelligence (AI) as an academic discipline, a milestone marked by the Dartmouth Conference of 1956. It was at this gathering that John McCarthy formally introduced the term "Artificial Intelligence," thus establishing AI as a distinct field of study and launching the Dartmouth Summer Research Project on Artificial Intelligence. The Dartmouth Project sought to explore various dimensions of machine intelligence, including language processing, concept formation, problem-solving techniques, and self-improvement mechanisms (McCarthy et al., 2006). The researchers at Dartmouth were cognizant of the technological limitations of their time, noting that "the speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, yet the major obstacle is not lack of machine capacity, but our inability to write programs that fully utilize what we have" (McCarthy et al., 2006). This insight echoed Alan Turing's earlier assertions that the potential of machine intelligence largely depends on the sophistication of its programming, while John von Neumann maintained a more cautious outlook on the prospect of replicating human-like intelligence (von Neumann, n.d.; Mühlenbein, 2009). The hypothesis that a significant portion of human thought involves manipulating words according to established rules of reasoning and conjecture anticipated the principles underlying modern large language models (LLMs), which generate and process text based on learned patterns and rules. Additionally, the Dartmouth team proposed that the difference between creative and routine thinking could involve "the injection of some randomness," a concept that has influenced contemporary AI strategies, including stochastic processes and probabilistic reasoning.

As the 1950s advanced, the foundational theories and expectations articulated during the Dartmouth Conference began to materialize in practical applications of artificial intelligence, culminating in one of the most significant developments: the Perceptron, devised by Frank Rosenblatt in 1958. Rosenblatt’s research was driven by fundamental inquiries into biological information processing, such as how information about the physical world is sensed, stored, and influences recognition and behavior (Rosenblatt, 1958). The Perceptron, a binary classifier capable of learning from input data, introduced adjustable weights that allowed it to classify spatial patterns into

predefined categories by assigning varying levels of importance to different inputs, mimicking the human brain's ability to prioritize stimuli (Nagy, 1991; Dreyfus, 1990). However, its limitations became apparent when Minsky and Papert's landmark work "Perceptrons" revealed that single layer perceptrons were inherently restricted in their pattern recognition capabilities.

As the practical applications of AI began to take shape, the development of ELIZA in 1966 by Joseph Weizenbaum at MIT represented a significant milestone in natural language processing. Weizenbaum describes ELIZA as "a program which makes natural language conversation with a computer possible" (Weizenbaum, 1966). The program's design was particularly innovative in its use of pattern matching and substitution methodology, which allowed it to engage in seemingly meaningful conversations despite its limited understanding. Weizenbaum noted that ELIZA's effectiveness stemmed from its ability to "maintain the illusion of understanding with so little machinery" (Weizenbaum, 1966). Building on Weizenbaum's insights into ELIZA's capabilities, the contemporary landscape of artificial intelligence, particularly with the advent of Large Language Models (LLMs), echoes similar philosophical debates and technical challenges. While modern LLMs like GPT-3 demonstrate an ability to generate responses that appear semantically rich, their operational underpinnings reveal that they do not possess conscious understanding. Instead, these models function through the manipulation of vast matrices containing weights and biases, which are adjusted during the training process to produce contextually appropriate language outputs.

This methodology resonates with Alan Turing's Imitation Game and John Searle's Chinese Room argument, both of which question the nature of understanding and intelligence in machines. Turing's test proposes that if a machine can mimic human behavior well enough to be indistinguishable from a human in conversation, it could be considered intelligent (Turing, 1950). Conversely, Searle's thought experiment argues that mere symbol manipulation (as performed by a computer) does not equate to genuine understanding or consciousness (Searle, 1980). In the case of modern LLMs, the debate continues as these models achieve greater linguistic fluency and complexity, prompting us to revisit what it means for a machine to "understand."

The paradox highlighted by these discussions is central to the design and application of LLMs. These models, though advanced, primarily excel in pattern recognition and statistical optimization, lacking an intrinsic understanding of the content they process or generate. This distinction is crucial in appreciating the capabilities and limitations of AI. While they can assist in automating and enhancing various tasks, including narrative generation in video games, their use must be critically evaluated to ensure they align with the intended human values and understanding.

As the focus of AI research transitioned from theoretical models to practical applications, the field experienced crucial developments that paved the way for today's sophisticated systems, including Large Language Models (LLMs) used in video game design and narrative generation. During the 1970s and 1980s, the development of expert systems marked a significant advance. These specialized information systems were designed to mimic the decision-making processes of human experts in specific domains, leveraging knowledge bases and inferential reasoning capabilities to address complex problems beyond the reach of conventional computer tools (Crouch, 1991). For instance, MYCIN is a renowned expert system that was developed to assist physicians in diagnosing and treating bacterial infections, particularly those affecting the blood and meningitis (McCarthy, 1984). The system used a set of if-then rules and a backward chaining inference engine to analyze patient data and provide diagnostic and treatment recommendations. MYCIN's knowledge base contained about 600 rules, which were derived from extensive interviews with infectious disease experts (Shortliffe, 1977).

Despite their utility, expert systems had inherent limitations. One of the key challenges was the knowledge acquisition bottleneck, which refers to the difficulty in eliciting and incorporating domain expertise into the system. The process of acquiring and representing expert knowledge can be time-consuming and complex, leading to delays in system development and deployment (Chu & Hwang, 2008). These systems were rule-based, relying heavily on predefined knowledge bases, which made them rigid and unable to learn from new data. This inflexibility highlighted the need for

systems that could not only apply existing knowledge but also adapt and learn from new experiences and information.

Inflexibility of expert systems highlighted the need for systems that could not only apply existing knowledge but also adapt and learn from new experiences and information. This need led to the development of backpropagation in the 1980s. Backpropagation, as described by Rumelhart, Hinton, and Williams (1986), is a learning algorithm that iteratively adjusts the weights in a neural network to minimize the error between the actual output and the desired output by propagating the error backward through the network's layers. The backpropagation algorithm involves two main phases: the forward pass and the backward pass. During the forward pass, input data is fed into the network, and the output is calculated through the network's layers using current weights. In the backward pass, the error between the predicted output and the actual output is calculated and propagated backward through the network to adjust the weights accordingly (Werbos, 1988). This advancement enabled networks to learn complex patterns and representations, setting the stage for future deep learning architectures.

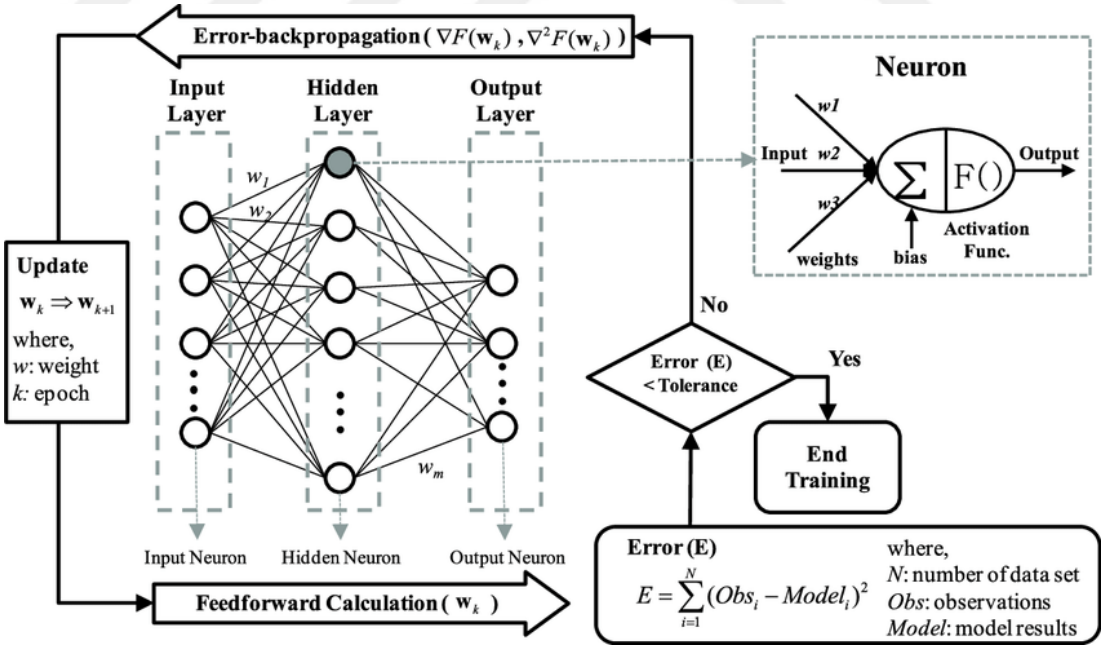

Figure 2 Diagram of feedforward neural network, illustrating the process of error-backpropagation and the update mechanism for weights during training.

Another significant key development is the introduction of Convolutional Neural Networks (CNNs) by Yann LeCun in the late 1980s. The study demonstrated the successful application of backpropagation in recognizing handwritten zip code digits (Lecun et al,1989). Convolutional Neural Networks (CNNs) have gained significant attention in the field of artificial intelligence due to their ability to perform feature extraction and classification within the network structure through learning, providing a certain degree of invariance while preserving the spatial topology of the input data (Lawrence et al., 1997). This development was crucial for application in areas ranging from automated image classification to advanced robotics, setting a new standard for how machines perceive and interpret visual information.

The other concept that becomes apparent after backpropagation is Recurrent Neural Networks (RNNs) are a type of artificial neural network that can retain memory of previous inputs through loops within the network architecture. This memory retention feature allows RNNs to effectively model sequential data and time-dependent relationships (Lawrence et al., 2000). RNNs have been shown to possess linguistic capabilities and can be used for tasks such as natural language grammatical inference (Lawrence et al., 2000). They are particularly useful for modeling syntactic processing and have been successfully applied in large-scale language modeling tasks (Emami & Jelinek, 2005).

However, traditional Recurrent Neural Networks (RNNs) face a well-known challenge called the vanishing gradient problem, which occurs when the loss function's gradients become extremely small as they are back propagated through the network during training, leading to slow or ineffective learning. To address this issue, Hochreiter and Schmidhuber (1997) introduced Long Short-Term Memory (LSTM) Networks, which incorporate a unique architecture with three key gates: an input gate, a forget gate, and an output gate. These gates regulate the flow of information by determining how much information should be memorized, discarded, or output at each step. This selective ability to remember or forget enables LSTMs to maintain relevant information in their memory, making them particularly powerful for capturing long-

term dependencies, especially in applications like natural language processing where understanding extended context is crucial.

The success of LSTMs in addressing the vanishing gradient problem marked a turning point that created a broader interest in neural networks. As neural networks evolved, the challenge of effectively training deep architectures became increasingly apparent, particularly due to issues like vanishing gradients and the computational complexity of backpropagation in networks with many layers. In response to these challenges, a pivotal development occurred with the introduction of a fast-learning algorithm for deep belief networks (DBNs) by Hinton, Osindero, and Teh (2006). This work revolutionized the training of deep networks by proposing a greedy layer-wise training approach, where each layer of the network is trained individually as a Restricted Boltzmann Machine (RBM) before being fine-tuned with a global optimization method. This study addressed the limitations of previous methods by significantly reducing the time required for training deep networks while also improving their performance on tasks such as image recognition. The techniques laid the groundwork for modern deep learning, making it feasible to train deep architectures that could learn complex hierarchical representations from large datasets and prepare the stages for large language models.

2.1.2 The rise of transformer architecture. Despite these successes, challenges remained in processing sequential data, particularly in tasks requiring understanding of long-range dependencies in text. To address these issues, the transformer architecture, which could be considered the foundation of all large language models, was introduced by Vaswani et al.(2017) in their seminal paper called "Attention Is All You Need". That architecture presents a self-attention mechanism that allows the model to weigh the importance of different parts of the input sequence when processing each element. One of the key advantages of the transformer architectures compared to RNNs and LSTMs, was its ability to parallelize computations which means transformers can process all elements of a sequence simultaneously. This parallelization increased computational efficiency which is allowing for the training of even larger models on more extensive dataset. This advancement echoes once again, the prescient observations from the 1956 Dartmouth Summer Project, where researchers noted that while computer capacity might be insufficient for simulating higher brain functions, the primary obstacle was not hardware limitations but the ability to fully utilize available resources. Another advantage of the architecture is the use of positional encoding. This technique allows the model to understand the order of elements in a sequence without relying on recurrence. By encoding the position of each token in the input sequence, transformers can maintain awareness of word order and relative positions which is essential for understanding the context and meaning in language.

The transformer architecture quickly became the foundation for state-of-the-art models in natural language processing, including BERT (Bidirectional Encoder Representations from Transformers) by Google (Devlin et al., 2018) and the GPT (Generative Pre-trained Transformer) series by OpenAI. These models demonstrated unprecedented performance on a wide range of language tasks, from sentiment analysis to question answering and text generation. The impact of transformers extended beyond natural language processing, influencing fields such as computer vision with models like the Vision Transformer (Dosovitskiy et al., 2020) and even game playing with the Decision Transformer for reinforcement learning (Chen et al., 2021). In the context of video game design and narrative generation, transformer-based

models have opened new possibilities for creating more dynamic, context-aware and coherent content via LLMs.

2.1.3 Large language models (LLMs): capabilities, limitations and applications. Large Language Models (LLMs) are neural network-based systems, which are backed by transformer architecture, trained on a vast corpara of text to perform a wide range of natural language processing tasks with their ability to generate human-like language and understand complex linguistic patterns (Fan et al., 2023; Hadi et al., 2024). They are distinguished by their ability to manage large datasets, including unstructured texts, and to discern semantic connections among words and phrases (Adnan & Akbar, 2019). This enables them to undertake tasks like translation, creative writing, and answering queries effectively (Hadi et al., 2024).

LLMs demonstrate versatility through their capability for few-shot and zero-shot learning, adapting to new tasks with little to no direct training (Brown et al., 2020). This adaptability, coupled with their contextual understanding facilitated by attention mechanisms, allows LLMs to generate coherent and contextually appropriate responses across diverse domains (Vaswani et al., 2017). Moreover, the principle of transfer learning underpins their efficiency, as pre-trained models can be fine-tuned for specific applications using relatively small datasets, significantly reducing the resources required for specialized tasks (Howard & Ruder, 2018).

As these models scale up, including billions of parameters, they display emergent capabilities like performing simple arithmetic and basic logic (Wei et al., 2022). This scalability, however, comes with challenges. The computational resources required for training and deploying LLMs raise concerns about energy consumption and accessibility (Strubell et al., 2019). Additionally, these models can unintentionally amplify biases present in their training data, necessitating careful ethical considerations in their development and application (Bender et al., 2021).

Despite their impressive capabilities, LLMs face significant challenges, particularly in terms of hallucination. Hallucination in LLMs refers to the generation of content that is unfaithful or nonsensical compared to the provided source or training

data (Ji et al., 2022). This issue is particularly concerning because LLMs generate highly fluent and convincing responses, making their hallucinations difficult to identify and potentially harmful in real-world applications (Ji et al., 2022).

Their adaptability of different domains, even with using small dataset to fine-tuning pre-trained models, makes the LLMs an inhibitor for humanity. In healthcare, for instance, LLMs have shown remarkable capabilities in medical education and clinical decision-making. As reported by Kung et al. (2023), ChatGPT, a prominent LLM, demonstrated performance comparable to or exceeding the passing threshold in the United States Medical Licensing Exam (USMLE), showcasing its potential as an interactive learning tool for medical students. Furthermore, in radiology, LLMs like ChatGPT have been explored for enhancing clinical workflow and promoting responsible utilization of radiology services (Rao et al., 2023).

In the financial sector, LLMs are making significant strides in tasks ranging from financial natural language processing to risk assessment and algorithmic trading. BloombergGPT, a 50 billion parameter language model trained on a diverse financial corpus, has revolutionized financial NLP tasks including news classification, entity recognition, and question answering (Wu et al., 2023).

The legal domain has also seen significant impact from LLMs. For example, GPT-4 scored in the 90th percentile on the Uniform Bar Examination, demonstrating its potential in legal practice (Katz et al., 2023). LLMs have been utilized to explore fiduciary obligations and assist in legal research, showcasing their ability to handle complex legal concepts and reasoning (Nay, 2023).

These diverse applications underscore the transformative potential of LLMs. To fully harness these capabilities, it is essential to develop specialized models tailored to specific domains while being mindful of their limitations. By advancing fine-tuning techniques and enhancing model robustness, some challenges, such as bias and hallucination, can be mitigated. As LLM technology evolves, continued research and interdisciplinary collaboration will be vital to ensuring that these powerful tools are

deployed responsibly and contribute positively to progress and innovation across various fields.

2.2 Quest

The word “quest” has deep linguistic roots dating back to Middle English which originated from the Latin “quaerere” meaning “to ask” or “to seek” (Oxford English Dictionary). This etymology reflects the essence of quests as pursuits or searches, often involving challenges or tasks to be completed. Quests are multifaceted elements in games and narratives, serving various purposes and carrying different meanings. Howard (2008) defines a quest broadly as "a journey" undertaken by a protagonist or player across a symbolic landscape, involving object collection, character interactions, and challenge overcoming to achieve a meaningful goal. In the context of gaming, Howard (2008) narrows this definition to a "goal-oriented activity" requiring players to navigate landscapes and complete challenges. Importantly, Howard (2008) emphasizes that quests bridge the theoretical gap between games and narratives, with the process of "discovering meaning" being a crucial aspect of what defines a quest. This meaning can manifest in three forms: as a reward for completing actions, as part of a larger narrative providing context or consequences, or as thematic representations of abstract ideas encoded in the quest's design elements. While Howard (2008) clarifies that quests are distinct from narratives, he acknowledges their close relationship, with narratives often providing motivation for quests and historical quest narratives influencing their implementation in video games. This intricate connection is evident in the popularity of quest-driven games like "King's Quest" and "The Elder Scrolls IV: Oblivion" (Howard, 2008).

Besides the meaning of quest, in academia there is no consensus on the definition of a quest which has significant importance in the context of procedural quest generation (Yu et al., 2020). Definition is important to create a set of rules that algorithms or LLMs will behave in the interspace of a defined frame. Especially in the context of LLMs and content generation, it is important to cluster the context gaps in a logical manner to allow models to have room to be creative.

Building on this understanding, Lima et al. (2019) proposed a more formalized representation of quests, defining them as a set of tasks to be accomplished by the player. They suggest that this set of tasks represents the plot or storyline of the quest. This definition is particularly useful in the context of procedural quest generation, as it provides a clear structure that can be algorithmically manipulated. Alpay (2022) proposed a comprehensive analytical framework to define quests, drawing inspiration from previous studies on quest definition. His framework centers around three core elements: Task, Meaning and Feedback.

Alpay's (2022) framework proposes that:

- Task refers to the specific actions the player needs to undertake within the game to progress through and complete the quest. It emphasizes that the task is distinct from the goal; the task encompasses the individual steps, while the goal signifies the overall desired outcome.
- Meaning delves into the significance and purpose behind the quest, extending beyond the mere execution of tasks. It identifies two types of meaning: Embedded meaning, which is intentionally incorporated by game developers for players to uncover, and Emergent meaning, which arises from players' interpretations and experiences that might not be explicitly dictated by the game's design.
- Feedback encompasses the game world's acknowledgment and response to the player's actions and completion of the quest. This can manifest as rewards or noticeable alterations within the game world. Alpay (2022) distinguishes between Interpretive feedback, which is subjective and dependent on the individual player's understanding, and Configurative feedback, which is deliberately designed by developers to provide specific outcomes based on player choices. The framework primarily analyzes Configurative feedback, as it is more objectively observable. Additionally, Reward is recognized as a subcategory of feedback, representing quantifiable benefits the player receives.

The proposed framework of Alpay has been used in this research to create a dataset for fine tuning the model.

2.3 Procedural Generation in Video Games

2.3.1 Overview. Procedural Content Generation (PCG) has become integral to video game development, particularly in addressing the growing demand for extensive and diverse game content (Värtinen et al., 2024). Togelius et al.(2011) define PCG as the 'algorithmic creation of game content with limited or indirect user input.' In the context of quest generation, PCG can be used to create varied and dynamic quest structures, potentially enhancing player engagement and game longevity. As games evolve towards larger open worlds and "Games as a Service" (GaaS) models, the traditional methods of manual content creation have become increasingly time-consuming and resource-intensive (Carli et al, 2011). PCG offers a solution by automating the content creation process, potentially leading to more diverse, engaging, and replayable games (Connor et al., 2017).

While PCG offers numerous benefits, its application in quest generation has faced challenges, especially with the rule-based approaches. Early attempts at PCG for quests, such as the work by Doran and Parberry (2011) using generative grammar, often relied on predefined rules, which could lead to predictable and repetitive quest structures (Värtinen,2022). Many PCG systems rely on templates or predefined structures, which is delimiting the system, leading to a lack of variety in generated quests. Players often perceive these quests as being limited to basic "transit/kill/use" patterns, lacking the depth and meaning found in hand-crafted quests (Harris & Caldwell ,2023) The challenge lies in balancing controllability with randomness, which has led to applications about using LLMs in PCG.

The probabilistic nature of LLMs comes from their ability to generate text by predicting the next token in a sequence based on the probability distribution of possible next words. This nature creates an opportunity to generate content that balances structure with randomness, compared to traditional rule-based systems. For instance, by adjusting the temperature parameter within LLMs, designers can control the degree of randomness in the generated content, allowing for the creation of quests that are both novel and contextually appropriate. This capability enables LLMs to move

beyond the limitations of rigid templates, introducing variability and creativity in quest design while maintaining coherence and alignment with the game's overarching narrative. This adaptability is particularly valuable in large, open-world games where a mix of structured and unpredictable content can enhance immersion and replayability.

2.3.2 Related work. The generation of quests through procedural methods has been a longstanding challenge in game AI, with research on the topic extending over nearly four decades (Kybartas & Bidarra, 2017). The earlier rule-based systems (Doran & Parberry, 2011) laid the groundwork for more advanced systems. For example, Choi et al. (2016) proposed a two-network model using RNNs to generate stories with correct grammar and context flow. Similarly, Ghosh et al. (2017) developed Affect-LM, an LSTM-based model incorporating affect categories and strengths. However, these RNN-based models have since been outperformed by transformer-based LLMs in various NLG tasks (Zhang et al., 2022).

Recent works have overcome some of the shortcomings through the use of language models for quest generation. Ammanabrolu et al. (2019) fine-tuned GPT-2 for creating quests in the form of cooking instructions in a text-based cooking game. Based on a small user study with 75 participants, they found that the GPT-2 quests were experienced as more valuable and coherent, but less surprising and novel than quests produced by random assignment or Markov chains.

Another notable example is the work by van Stegeren and Myśliwiec (2021), who fine-tuned GPT-2 for generating quest descriptions from the perspective of NPCs in World of Warcraft. Their approach used a dataset of 24,981 quests from World of Warcraft, annotated with special tags to denote the quest title, objective, and description. By fine-tuning GPT-2 on this structured data, they were able to generate new quest descriptions given a title and objective as input. Van Stegeren and Myśliwiec (2021) conducted a user study with 32 participants to evaluate the generated quest descriptions against human-authored ones. While the generated descriptions scored lower on average for language quality, coherence, and novelty, the differences in surprise and creativity ratings were not statistically significant. Importantly, their

results showed that the fine-tuned model could sometimes produce outputs on par with or exceeding human-written quest descriptions.

Furthermore, research by Todd et al. (2023) has shown promising results in using LLMs for level generation in games. They fine-tuned GPT-2 on a dataset of Sokoban levels and found that the model could generate novel, playable puzzles. Interestingly, their results indicated that while GPT-2 struggled with limited training data, larger models like GPT-3 were better able to generalize from smaller datasets, suggesting that more advanced LLMs could potentially overcome some of the limitations faced by earlier models in quest and content generation.

Värtinen et al. (2022) conducted a more comprehensive study using GPT-2 for quest generation in RPGs. They created a novel dataset of 978 quests from six different RPGs and fine-tuned GPT-2, resulting in a model called Quest-GPT-2. Their method involved replacing proper nouns and numbers with placeholder tokens to reduce variance during training. In a large-scale user study with 349 participants rating 500 quest descriptions, they found that approximately 20% of the generated descriptions were deemed acceptable by human critics. Although this indicates progress, the authors highlighted significant variation in quality across quests and persistent challenges with entity distinction and logical coherence. They also explored the potential of GPT-3, noting promising improvements in coherence compared to GPT-2. Värtinen et al. (2022) proposed several applications for Quest-GPT-2, including its use as an assistant for co-creative quest writing, generating quest ideas, and offline generation with human curation, underscoring the evolving role of LLMs in game content creation.

Mishra (2023) demonstrated the efficacy of using knowledge graphs to model narrative worlds in quest generation, providing contextual data that significantly enhances the quality of generated quests using the data from the research of Värtinen et al. (2022). By embedding knowledge graphs into LLMs, Mishra (2023) was able to capture the intricate relationships between narrative entities, leading to quests that were more coherent and aligned with the story's plot points. This approach shows

promise in addressing common challenges such as entity distinction and logical coherence, which have been persistent issues in earlier models (Mishra, 2023).

These studies show the real potential of use of LLMs in the context of game design and development. Those researches are demonstrating the importance of fine-tuning LLM based approaches, and the importance of structured datasets to improve the quality of generated contents. As the LLMs get bigger and more efficient, those possibilities will expand.

2.4 Fine-Tuning Large Language Models (LLMs)

2.4.1 Introduction to fine-tuning. Fine-tuning plays an essential role in tailoring Large Language Models (LLMs) to perform specific tasks by training them on task-focused datasets. This process, rooted in the principles of transfer learning, enables a pre-trained model—initially trained on a broad and diverse dataset—to be further refined on a smaller, domain-specific dataset relevant to the target application (Howard & Ruder, 2018). The value of fine-tuning lies in its ability to transform a general-purpose model into one that excels in a particular domain, such as generating contextually appropriate quests in video games. This approach not only enhances the model's performance on specialized tasks but also significantly reduces the computational resources and time required compared to training a new model from scratch. The versatility provided by fine-tuning makes LLMs powerful tools across various fields, including narrative generation and healthcare, where domain-specific knowledge and tailored responses are paramount (Ruder, 2019).

2.4.2 Techniques for fine-tuning LLMs. Building on the foundational importance of fine-tuning, this section explores the diverse techniques employed to optimize LLMs for specific tasks, especially in creative contexts such as video game quest generation. Each technique offers unique advantages and is suited to different scenarios, depending on the computational resources available and the complexity of the task.

Full Fine-Tuning involves updating all the parameters of a pre-trained model on a new, task-specific dataset. This comprehensive approach allows the model to integrate the specific nuances and characteristics of the new task deeply. While this method can lead to highly specialized outputs, it is computationally intensive and requires significant resources. In a study by Latif and Zhai (2023), full fine-tuning of GPT-3.5-turbo was used to enhance the model's performance in automatically scoring student responses in educational settings. By training on a diverse dataset of student responses and expert evaluations, the fine-tuned model outperformed BERT, particularly in tasks requiring nuanced understanding and context-specific judgments (Latif & Zhai, 2023). This approach is particularly effective in creative tasks like quest generation in video games, where the model must understand and generate content that aligns with complex narrative structures and player expectations.

In contrast, Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged as a more resource-conscious alternative to full fine-tuning. PEFT strategies, such as Adapter modules and Low-Rank Adaptation (LoRA), focus on modifying only a small subset of a model's parameters. By selectively adjusting these parameters, PEFT methods achieve task-specific adaptation while preserving the bulk of the pre-trained model's knowledge. This approach not only reduces the computational burden but also mitigates the risk of catastrophic forgetting, where the model loses previously learned information while acquiring new knowledge (Houlsby et al., 2019; Hu et al., 2022). Additionally, quantization methods have been introduced to further optimize the fine-tuning process by reducing memory and computational requirements. Quantization involves lowering the bit precision used to represent model parameters, enabling LLMs to operate efficiently on lower-power hardware. Techniques such as QLoRA combine quantization with low-rank adaptation, facilitating effective fine-tuning even on consumer-grade devices (Mishra, 2023). This innovation not only makes sophisticated AI models more accessible but also broadens their applicability across diverse fields, as it allows high-performance models to be utilized in environments with limited computational resources (Dettmers et al., 2023).

2.5 Evaluation of AI-Generated Content

Evaluating AI-generated content, particularly for creative domains like video game quest design, poses unique challenges. While traditional natural language processing (NLP) evaluation metrics can provide some insights, they often fall short of capturing the nuanced and subjective qualities that make a quest engaging, coherent, and contextually appropriate. Consequently, a combination of quantitative and qualitative approaches is essential to comprehensively assess the effectiveness of AI-generated content. Therefore, a comprehensive evaluation strategy should combine quantitative metrics with qualitative methods to gain a holistic understanding of the model's performance.

2.5.1 Quantitative evaluation metrics. Quantitative metrics provide objective, reproducible measures of AI performance, but their application to creative tasks requires careful consideration.

2.5.1.1 BLUE (Bilingual evaluation understudy). The BLEU metric is widely used in natural language processing to evaluate machine-generated text by comparing it against a set of reference texts. It scores the overlap between the generated text and the reference texts, offering a measure of similarity (Papineni et al., 2002). However, in creative domains like quest generation, BLEU's reliance on surface-level similarity can be a limitation, as it tends to favor outputs that closely resemble the reference texts and may penalize more innovative or creative expressions. This limitation is particularly relevant in the context of video game design, where creativity and variability are crucial for player engagement.

2.5.1.2 ROUGE (Recall-oriented understudy for gisting evaluation). ROUGE metrics evaluate the overlap of n-grams, word sequences, and word pairs between generated text and reference texts, making them useful for tasks like summarization and content generation (Lin, 2004). Although commonly applied in summarization, ROUGE can also assess the coherence and relevance of AI-generated quests. However, similar to BLEU, ROUGE may not fully capture the creative and thematic aspects of quests, focusing instead on textual similarity.

2.5.1.3 METEOR (Metrics for evaluation of translation with explicit ORdering). METEOR is a more sophisticated metric that accounts for synonymy and stemming, making it more closely aligned with human judgment compared to metrics like BLEU or ROUGE (Banerjee & Lavie, 2005). By recognizing similar meanings conveyed through different words and forms, METEOR offers a more nuanced evaluation of linguistic variability and semantic fidelity in generated quests. This capability is particularly valuable in assessing the quality of AI-generated content, as it captures subtle differences in word choice and phrasing, allowing for a more comprehensive assessment of whether the content effectively conveys the intended meaning and context.

2.5.1.4 BERTScore. BERTScore is a relatively recent metric that leverages contextual embeddings from pre-trained transformer models like BERT to assess the semantic similarity between generated text and reference texts, providing a more nuanced evaluation compared to traditional metrics such as BLEU and ROUGE (Zhang et al., 2020). Unlike these older metrics, which often focus on surface-level text overlaps, BERTScore captures deeper semantic relationships, making it particularly suitable for evaluating narrative content and quest generation in video games. This is crucial in contexts where the meaning, coherence, and thematic consistency of the text are more important than exact word matching.

2.5.1.5 Perplexity. Perplexity is a common metric used to evaluate the performance of language models, particularly in the context of sequence prediction tasks. It measures how well a model predicts a sequence of words, with a lower perplexity indicating a model's better performance at predicting the next word in a sequence (Jelinek et al., 1977). While perplexity is an important indicator of a model's linguistic fluency and coherence, it does not necessarily correlate with the overall quality or creativity of the generated content. In the context of quest generation, lower perplexity scores may suggest that the generated quests are more fluent and contextually appropriate, yet they do not provide insights into the novelty or engagement of the quests.

2.5.1.6 F1 Score. The F1 score, which balances precision and recall, is a critical metric for evaluating AI-generated content, particularly in scenarios where precision (the relevance of generated content) and recall (the coverage of necessary content elements) are both important (Sasaki, 2007). It represents the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. In the context of video game quest generation, the F1 score can be used to assess how well the AI-generated quests align with the desired narrative elements, ensuring that the content is both accurate and comprehensive. By balancing precision and recall, the F1 score offers a nuanced evaluation of the AI's ability to generate relevant and complete content, making it a valuable tool for identifying areas where the model may need further refinement.

2.5.2 Qualitative evaluation methods. Qualitative evaluation methods are crucial for assessing the nuanced and subjective qualities of AI-generated content, particularly in creative tasks such as video game quest design. These methods go beyond numerical metrics, focusing on the richness, depth, and contextual appropriateness of the generated content, offering insights into how well AI-generated quests resonate with human players and align with narrative themes.

2.5.2.1 Human evaluation. Human evaluation is a way to evaluate the quality and accuracy of model-generated results through human participation (Chang et al.,2023). Earlier research by Novikova et al. (2017) shows that automatic evaluation metrics for NLG systems often fail to accurately mirror human ratings, highlighting the necessity for human evaluations. Human evaluators can provide nuanced judgments that go beyond what automatic metrics can capture, particularly in tasks involving creativity, coherence, and contextual relevance. For example, human evaluation is essential in assessing the fluency and naturalness of generated text, as well as the alignment of the content with user expectations and cultural norms.

The evaluation of Large Language Models (LLMs) through human evaluation is a complex process that requires careful consideration of multiple factors to ensure reliable and accurate results. Fu et al. (2023) provide a comprehensive overview of these critical aspects in their survey on LLM evaluation. They emphasize three key elements: the number of evaluators, evaluation criteria, and evaluator expertise. The authors stress the importance of selecting an appropriate number of evaluators to ensure adequate representation and statistical significance, which contributes to a more nuanced understanding of the LLMs being evaluated. Building on the "3H rule" (Helpfulness, Honesty, and Harmlessness) proposed by Askell et al. (2021), Fu et al. (2023) expand the evaluation criteria to include six essential components: accuracy, relevance, fluency, transparency, safety, and human alignment. Each of these criteria serves a specific purpose in assessing different aspects of LLM performance, from the factual correctness of the generated text to its ethical implications and alignment with human values. Additionally, the authors highlight the importance of evaluator expertise, emphasizing that evaluators should possess relevant domain knowledge, task familiarity, and methodological training to accurately assess domain-specific text generated by LLMs.

However, human evaluation faces several challenges, including reproducibility issues, potential inconsistencies, and high costs in terms of time and resources (Clark et al., 2021; Karpinska et al., 2021). Also, lack of reproducibility remains as a limitation. In light of the limitations associated with traditional human evaluation methods, recent research has explored the potential of using large language models

(LLMs) as expert reviewers in evaluating AI-generated content. Chiang and Lee (2023) demonstrate that LLMs can produce evaluations consistent with those of human experts, particularly in tasks like open-ended story generation, where creativity and contextual relevance are crucial. Their study compared LLM-based evaluation against expert human evaluation on tasks such as open-ended story generation and adversarial text attacks. The authors found that advanced LLMs like GPT-3.5 (specifically the InstructGPT model text-davinci-003) and ChatGPT were able to rate the quality of generated stories in ways that closely aligned with expert human evaluators. For instance, both LLMs and human experts consistently rated human-written stories higher than AI-generated stories across attributes such as grammaticality, coherence, likability, and relevance to prompts. This approach offers several advantages, including improved reproducibility, independence between individual sample evaluations, lower costs, and faster turnaround times compared to traditional human evaluation methods (Chiang & Lee, 2023) However, the use of LLMs as evaluators is not without its own challenges, particularly regarding the potential for bias and the lack of nuanced understanding that human reviewers provide. Despite these challenges, integrating LLM-based evaluations with human assessments could offer a more comprehensive approach to evaluating AI-generated content, particularly in creative domains like video game quest design.

2.5.2.2 Expert review. Expert review plays a crucial role in the qualitative evaluation of AI-generated content, particularly in domains where creativity, thematic coherence, and adherence to narrative structures are paramount. Unlike general human evaluations, which may lack the depth of understanding required to fully assess the intricacies of game design, expert reviews draw upon the deep knowledge and experience of professionals who are well-versed in the nuances of the field. These experts are often provided with a structured evaluation framework to ensure that their assessments are systematic and focused on key aspects of game design.

Moreover, expert review is not only about evaluating the end product but also about understanding the process by which the AI model arrived at its output. Experts can provide insights into whether the model's approach to quest generation is consistent with best practices in game design, offering recommendations for improving the

model's logic, creativity, and adherence to thematic consistency. This iterative feedback loop between experts and AI models is essential for advancing the state of AI-generated content, particularly in creative fields where qualitative nuances are often more important than quantitative metrics.

2.5.2.3 Comparative analysis. Comparative analysis is a powerful method for evaluating the performance of AI-generated quests by benchmarking them against established standards or human-created content. This approach provides a direct way to assess the quality and effectiveness of AI models, particularly in creative domains where the bar for success is set by human expertise.

The comparative analysis is not limited to qualitative assessments; it also incorporates quantitative metrics such as BLEU, ROUGE, and BERTScore to provide a comprehensive evaluation. By systematically comparing the outputs of fine-tuned and non-fine-tuned models, this research aims to demonstrate the value of fine-tuning LLMs with synthetic data in producing high-quality, contextually appropriate quests.

Chapter 3

Methodology

3.1 Research Questions

This research seeks to answer the question of “How does fine-tuning Large Language Models (LLMs) with synthetic data derived from structured quest information affect the quality and diversity of quests generated for video games?”. Sub-questions are as follows:

- How can data from community-driven platforms, like the Assassin's Creed Wiki, be effectively structured and utilized to enhance LLM-generated quests?
- In what ways can fine-tuning LLMs innovate traditional narrative techniques in video games, particularly in terms of integrating interdisciplinary approaches and evolving narrative structures?
- How can the quality of AI-generated content be improved by refining the data collection and structuring process, and what role do hyperparameters play in optimizing these outputs?
- What roles do hyperparameters play in the fine-tuning process, and how do they affect the output quality of LLM-generated content?
- How do system prompts influence the effectiveness of both base and fine-tuned models in generating high-quality quests?

3.2 Purpose of the Study

The purpose of this study is to explore the applications of fine-tuning Large Language Models (LLMs) using synthetic data derived from structured quest information in the context of video game quest design, with a specific focus on Assassin's Creed Odyssey. This study aims to:

- Develop and evaluate a methodology for fine-tuning LLMs to generate contextually appropriate quests that align with the game.
- Assess the effectiveness of fine-tuned LLMs in producing high-quality quest content compared to non-fine-tuned models.
- Examine the influence of hyperparameters and system prompts on the performance of both base and fine-tuned models, to optimize the quality of AI-generated content in video games.
- Investigate the potential of integrating AI-generated content into the creative processes of game design.

3.3 Significance of the Study

This research is significant in its advancement of AI applications, particularly by fine-tuning Large Language Models (LLMs) with domain-specific data, thus contributing to the expanding body of knowledge on how AI can be effectively leveraged in creative industries, such as video game development. It introduces innovative methodologies for quest design, offering game designers advanced tools to enhance and streamline the creation of engaging and complex narrative content. Situated at the intersection of game design, narrative theory, and artificial intelligence, this study provides valuable interdisciplinary insights and tools, demonstrating how collaboration between different disciplines can drive innovation and improve creative processes in game development. The practical implications of this research are substantial, as the findings have the potential to inform the development of AI-assisted tools that could significantly reduce the time and resources required to create immersive and diverse game content, thus benefiting the broader gaming industry.

3.4 Research Design

This study adopts a mixed-methods approach, combining quantitative data analysis with qualitative narrative analysis to explore the impact of fine-tuning Large Language Models (LLMs) on the generation of quests in video games, specifically within the context of Assassin's Creed Odyssey. The research is grounded in a comprehensive dataset derived from the quests documented in the Assassin's Creed

Wiki, a community-driven platform that offers a wealth of structured and unstructured data. This dataset includes critical information on quests, such as narrative components, player choices, and outcomes, providing a robust foundation for the study.

The data was carefully structured to ensure consistency and manageability. Custom tools were employed to convert the original XML files into JSON format, enabling a more streamlined analysis. This structured data comprises essential quest attributes, including names, types, locations, and detailed dialogue segments. During this process, minor typographical errors were corrected, and uniformity in data representation was maintained to ensure the integrity of the dataset.

A subset of this structured data was then utilized to generate synthetic dataset for fine-tuning the LLM. This synthetic data was created by applying a quest framework that is inspired by the work of Alpay (2022), that incorporates various task types, narrative meanings, and feedback loops. The goal was to expose the LLM to a wide array of contextually rich quest examples, thereby enhancing its ability to generate contextually appropriate and narratively coherent quests.

The study selected both the ChatGPT 3.5 Turbo and ChatGPT 4o-mini models as the foundational LLMs for fine-tuning, recognizing their respective strengths in natural language understanding and generation. The fine-tuning process for both models involved inputting synthetic data, guided by carefully crafted system instructions aimed at optimizing their performance. These instructions were designed to ensure that the generated quests were relevant and coherent within the game's narrative framework. Throughout the process, various hyperparameters were adjusted, and multiple iterations were conducted to refine the outputs of both models.

The study employed both quantitative and qualitative methods to evaluate the fine-tuned LLM's output. Quantitative analysis focused on metrics such as the diversity of quest types generated, the accuracy of narrative sequencing, and the contextual relevance of the quests to the game's overarching narrative. On the other hand, qualitative narrative analysis was used to assess the depth and richness of the

generated quests. This involved examining the narrative components, such as character development, moral dilemmas, and thematic references, to determine how well the fine-tuned model integrated these elements into the quests and how it enhanced the overall gaming experience.

3.5 Game Selection

Assassin's Creed Odyssey was selected for this study due to its complex quest architecture, intricate narrative elements, and expansive open-world environment, offering an ideal dataset for fine-tuning Large Language Models (LLMs). Its diverse array of quests, encompassing detailed storylines, multifaceted character interactions, and player-driven choices, provides a comprehensive platform for exploring how LLMs can generate contextually appropriate and engaging content. The game's historical and cultural setting adds another layer of complexity, challenging LLMs to produce narratives that are both coherent and aligned with historical contexts. Additionally, the wealth of community-contributed resources, such as the Assassin's Creed Wiki, facilitates the extraction and structuring of quest data, making it easier to create synthetic datasets for model fine-tuning. By focusing on a game that exemplifies modern video game narrative and structural complexities, this study aims to generate findings with broad implications for contemporary game development, offering insights into the integration of AI-assisted tools in creative processes within the gaming industry.

3.6 Data Collection

3.6.1 Manual data collection from gameplay. The manual data collection involved playing through 390 quests in Assassin's Creed Odyssey to systematically document essential quest information. Personal engagement with the game allowed the researcher to capture key attributes such as Chapter Title, Chapter Name, Quest Type, Quest Name, and Location, which were recorded in an Excel file. This process allowed for a thorough understanding of the game's quest structure, ensuring the data collected was accurate and comprehensive, thus providing a solid foundation for subsequent analysis.

3.6.2 Data source. The primary data for this study was collected from the Assassin's Creed Wiki, a comprehensive repository of Assassin's Creed games content documented by voluntary players. The wiki provides the functionality to download pages within specific categories in XML format. Specifically, the dataset was extracted from the category "Memories using the Animus HR-8.5" which contains information blocks about quests appearing in both Assassin's Creed: Odyssey and Assassin's Creed: Valhalla. The data was accessed under the CC-BY-SA license, which permits adaptation and redistribution, provided that proper attribution is given, and derivative works are shared under the same or similar terms.

3.7 Data Processing and Structuring

The data processing and structuring phase was a pivotal component of this research, involving the transformation of raw, community-contributed data into a structured and reliable dataset suitable for fine-tuning Large Language Models (LLMs). This process addressed the challenges inherent in working with community-sourced content, particularly from the *Assassin's Creed* Wiki, which was supplemented by personal gameplay documentation. To ensure data integrity and address these challenges, a multi-stage processing pipeline was developed, supported by a series of custom tools. These tools were designed with flexibility and reusability in mind, making them valuable not only for this study but also for future research

involving other games in the *Assassin's Creed* series or similar narrative-driven games that rely on Fandom Wiki websites.

3.7.1 XMLParser for JSON conversion. The conversion of XML files to JSON format enabled the transformation of raw, community-contributed content into a structured format that could be readily utilized for further analysis and fine-tuning of Large Language Models (LLMs). This process, facilitated by the custom-developed XMLParser tool, involved extracting and organizing data from the XML files sourced from the Assassin's Creed Wiki.

The choice of JSON as the target format was driven by its inherent advantages over XML in terms of flexibility, readability, and ease of use in modern data processing environments. JSON's lightweight and hierarchical structure made it particularly well-suited for handling the complex, nested data typical of video game quests, such as branching dialogues and multifaceted narrative sequences. Additionally, JSON's compatibility with a wide range of data manipulation libraries allowed for seamless integration into subsequent stages of the research pipeline.

During the conversion process, the XMLParser extracted key elements from the XML data, such as quest names, descriptions, and associated metadata, transforming them into a more accessible JSON structure. This step streamlined the data for subsequent processing and included preliminary data cleaning, where minor typographical errors were corrected, and the data was standardized to ensure consistency.

The resulting JSON dataset preserved the narrative and structural complexities of the original quest data, providing a reliable foundation for further processing, including data cleaning, structuring, and fine-tuning of LLMs, thereby supporting accurate analysis and generation of complex narrative content.

3.7.2 Data understanding. After converting the XML data to JSON format, the next crucial step was to gain a comprehensive understanding of the data structure. The JSON data, derived from the Assassin's Creed Wiki, comprised various information blocks related to quests in the game. Each quest was represented as an object with multiple key-value pairs, where the keys corresponded to different attributes of the quest, and the values provided the actual data. Here are some of the critical keys identified:

- Quest Name: The title of the specific quest or mission in the game.
- Section Dialogue: A comprehensive transcript of conversations that occur during the quest. This includes character speech, player dialogue options, and sometimes descriptions of character actions or emotions during the conversation. It provides a detailed account of the story's progression and character interactions.
- Section Outcome: The results or consequences of completing the quest.
- Section Trivia: Interesting facts about the quest or related game elements.
- Section Behind the Scenes: Information about the development process, creative decisions, or interesting facts about how this quest or element was created. This might include designer intentions, historical research done for the quest, or connections to real-world events or mythology.

3.7.3 Tool development for data manipulation. The process of refining raw quest data into a structured and analyzable format necessitated the development of several custom tools. These tools were created to efficiently manage the large dataset extracted from the Assassin's Creed Wiki, facilitating operations such as data cleaning, dialogue structuring, and hierarchical organization. Each tool was specifically tailored to address challenges encountered during data processing, ensuring that the dataset was optimized for subsequent fine-tuning of Large Language Models (LLMs).

3.7.3.1 Data manipulator. The *DataManipulator* tool was a cornerstone of the data processing pipeline, designed to handle a wide range of data manipulation tasks, including loading, retrieving, and structuring quest information. This tool provided the researcher with the flexibility to interact dynamically with the dataset. Key features of the *DataManipulator* include:

- Loading and Saving Data: The tool loads data from JSON files and saves modifications back to JSON, maintaining data integrity and facilitating easy access throughout the research process.
- Data Cleaning: Functions to remove null values, sanitize filenames, and replace empty strings help maintain the cleanliness and uniformity of the data set.
- Data Retrieval and Manipulation: The tool allows for the retrieval of specific quests or dialogues by index or name and offers filtering based on criteria such as appearance or quest name.
- Data Structuring: The tool can categorize quests, match them with files in specified folders based on chapter types, and process structured dialogues, which is crucial for managing large datasets effectively.

This tool is vital for managing the extensive quest data and simplifies many operations that would otherwise be manual and time-consuming, thereby enhancing research efficiency.

3.7.3.2 Dialogue data structurer. The *DialogueDataStructurer* tool was developed to further refine quest data, particularly in understanding the complex branching dialogues typical of narrative-driven games. This tool employed a set of regular expressions to parse and segment dialogue text into distinct components, such as player choices, narrative sequences, and conditions. Each segment was assigned a unique identifier, enabling precise tracking of dialogue flow and interaction points within the quests.

The *DialogueDataStructurer* was particularly effective in identifying and structuring nested dialogues, common in quests involving multiple layers of decision-making and branching outcomes. By deconstructing dialogues into their core

components, the tool facilitated a deeper analysis of how narrative elements interact and evolve within the game's quest framework. This structured dialogue data was then integrated into the larger dataset, ensuring consistency and clarity in how quests were represented. While this module was not used directly in the core research, it provided significant value for further analysis, offering ease of use for practitioners and researchers interested in exploring intricate narrative dynamics.

3.7.4 Categorization and clustering. The categorization and clustering of quests was a critical process in understanding the complex narrative structure of Assassin's Creed Odyssey. This process involved meticulous manual curation, informed by direct gameplay experience and supplemented by online resources, followed by automated matching using the custom DataManipulator tool.

Each quest was exported as an individual JSON file, named after the quest, itself. These files were manually sorted into a hierarchical folder structure that reflected the game's diverse chapter organization. The dataset included quests beyond the 390 directly played by the researcher, with additional quests categorized based on reputable online sources to ensure comprehensive coverage.

The resulting folder structure represented the main chapter types, including Odyssey Chapter (main storyline quests), World (general side quests), Character (character-specific side quests), DLC Chapters, and The Lost Tales of Greece.

The sequencing system was tailored to reflect the game's varying narrative structures across different chapter types. For Odyssey Chapters, these main storyline chapters were the only ones assigned a Chapter_SequenceID, reflecting their linear progression in the main narrative. Within each Odyssey Chapter, quests were sequenced numerically to indicate their order. For Character and The Lost Tales of Greece Chapters, while these chapter types did not have an overall sequence, they contained various sub-chapters. Within each sub-chapter, quests were sequenced to reflect their narrative order. World and DLC Chapters were not assigned sequence IDs, reflecting their more open-ended nature in the game.

Following the manual categorization, the DataManipulator tool was employed to ensure accurate association of the structured quest data with the manually created folder structure. The tool scanned the folder structure, identifying chapter types, chapter names, and where applicable, sub-chapters and quest sequences. Using fuzzy string-matching techniques, the tool associated each quest in the JSON data with its corresponding file in the folder structure. Based on the matching results, the tool assigned metadata to each quest entry, including Chapter Type and Chapter Name for all quests. Odyssey Chapter quests were assigned an additional Chapter Sequence ID, while Character and Lost Tales quests were assigned sub-chapter names and quest sequence within sub-chapters. World and DLC quests did not receive additional sequencing metadata.

The automated matching results were carefully reviewed to ensure accuracy, with any discrepancies addressed individually. This nuanced approach to categorization and sequencing resulted in a comprehensive, structured dataset that accurately represented the game's complex narrative organization. It preserved the linear progression of the main story while also capturing the more flexible structures of side content and DLCs. This organized dataset formed the foundation for subsequent analysis and LLM fine-tuning, enabling a deep exploration of Assassin's Creed Odyssey's multifaceted quest design and narrative structure, including both its linear and non-linear storytelling elements.

3.7.5 Data filtering, cleaning, flattening for final data. The initial XML file was sourced from the "Memories using the Animus HR-8.5" category, which included information from both Assassin's Creed: Odyssey and Assassin's Creed: Valhalla. Since this research is specifically focused on Assassin's Creed: Odyssey, quests related to other games were filtered out using the Data Manipulator tool.

Additionally, the dataset included irrelevant key-value pairs, such as tags related to website maintenance and other metadata pertinent to the contributors who documented the quest pages. To enhance the dataset's clarity and reduce its size, these unnecessary elements were systematically removed.

Moreover, quests that lacked the "Section Dialogue" key were excluded from the dataset. This decision was made because the "Section Dialogue" key was crucial for this study, as it contained essential details of quest design, including dialogues, narratives, choices, and other elements critical to the research objectives.

Finally, to further streamline the dataset and enhance its usability, the nested key-value pairs were flattened. This process involved restructuring the data so that each quest was represented by a single level of key-value pairs, thereby improving the dataset's clarity and accessibility for subsequent analysis.

3.8 Fine-tuning

The fine-tuning of the ChatGPT 3.5 Turbo and ChatGPT 4o-mini models sits in the center of this study, aiming to explore the model's ability to generate complex, contextually accurate, and narratively engaging quests in video games. This process involves refining the models' output by adjusting their internal parameters using a specialized dataset. Each dataset entry consists of a system prompt, a user query, and a structured response generated by the assistant. The fine-tuning process aligns the models with the desired quest structure, depth, and narrative coherence. A carefully designed pipeline, developed specifically for this project, processes the structured quest data from previous stages into this format, facilitating effective fine-tuning.

3.8.1 Synthetic dataset creation pipeline for fine-tuning. The creation of a synthetic dataset is a multi-layered process, designed to convert structured quest data into a format suitable for fine-tuning LLMs. The dataset was constructed using JSON data previously developed in this study, which contains detailed quest goals, narrative explanations, and dialogues. Given the vast amount of text data, managing token count was crucial to optimizing the cost-effectiveness of fine-tuning.

Building upon Alpay's (2022) structured quest design framework, a specialized instruction prompt was developed to transform the JSON data into a format ideal for LLM training. This pipeline begins with the ingestion of structured JSON data representing individual quests, which are processed through the Gemini Pro 1.5 model.

The model simulates the response a quest designer might expect from a conversational AI, incorporating complex elements such as player objectives, inter-character relationships, historical context, and environmental interactions.

The pipeline starts by ingesting structured JSON data representing individual quests, which are then processed through the Gemini Pro 1.5 model. This model generates responses that align with what a game designer might expect from a conversational AI, integrating complex elements such as player objectives, inter-character relationships, historical context, and environmental interactions to create a dynamic narrative and gameplay interplay. To manage this effectively, the process was conducted on the Google AI Studio platform, where custom instructions guided the model's output. These instructions were essential in ensuring that the quest descriptions adhered to the desired structure and complexity, focusing on narrative intricacies, task integration, feedback loops, and environmental interactions. Narrative progression was tracked using identifiers like Chapter_SequenceID and Quest_SequenceID, which maintained coherence without explicit mention in the output, thus preserving the game's immersive experience while enriching the dataset.

Once the initial assistant reply was generated, it was re-entered into the Gemini Pro 1.5 model with a carefully crafted prompt simulating user input that could have led to the specific assistant reply. This iterative process—generating a reply and then crafting a corresponding input—resulted in a dataset that closely mimics real-world user interactions.

These system-user-assistant pairs were then compiled into a structured dataset, providing the necessary examples to fine-tune the LLMs. The fine-tuning process utilized this dataset to train the models, ensuring they could accurately replicate the detailed quest analysis and narrative structuring demonstrated in the example outputs.

3.8.2 Fine-tuning procedure. The fine-tuning procedure involved training the ChatGPT 3.5 Turbo and ChatGPT 4o-mini models using the synthetic dataset created in the previous step. The fine-tuning process aimed to improve the models' ability to generate high-quality, contextually appropriate quest content by iteratively adjusting various hyperparameters.

Several hyperparameters were tested during fine-tuning, including learning rate, batch size, and the number of training epochs. The learning rate was carefully calibrated to balance the speed of learning with the risk of overfitting, ensuring that the models could generalize well to new quest scenarios. Batch size was adjusted to optimize computational efficiency while maintaining model stability during training. The number of training epochs was determined based on the model's performance on validation data, allowing for sufficient training without overfitting.

Throughout the fine-tuning process, the models were iteratively adjusted with different hyperparameters. Although these metrics might not be ideal for evaluating creative tasks, key model metrics such as train accuracy, valid mean token accuracy, and valid loss were monitored to gain insights into how the model was learning and producing outputs. These metrics provided a quantitative foundation for understanding the model's behavior, guiding the adjustments to hyperparameters.

In addition to these metrics, an expert review was conducted during the fine-tuning process. This review involved assessing the generated quests to evaluate their narrative coherence, diversity of quest types, and contextual relevance to the game's overarching narrative. By combining both quantitative metrics and expert qualitative assessments, the aim was to achieve the best possible balance between model performance and the creative demands of quest generation.

3.9 Quest Generation

In this study, a systematic approach was employed to generate 720 unique quests for Assassin's Creed Odyssey by combining 15 distinct user input prompts, 4 different system role prompts, and 12 models—comprising both base models like GPT-4o Mini and ChatGPT 3.5 Turbo, as well as their fine-tuned counterparts.

3.9.1 User input prompts. The 15 user input prompts were designed to challenge the LLMs with intricate narrative concepts, ensuring the generated quests were rich in thematic depth and aligned with the historical and philosophical underpinnings of *Assassin's Creed Odyssey*. These prompts ranged from exploring existential themes like the Ship of Theseus paradox to the ethical implications of rewriting history, ensuring a diverse set of quests that reflect the game's narrative complexity.

3.9.2 System prompts. To guide the LLMs in generating quests that are contextually appropriate and thematically consistent, four different system role prompts were utilized. These prompts provided varying degrees of guidance and details to the models which influencing the tone, structure, and gameplay elements of generated quests:

1. System Prompt 0 (Empty Prompt): This served as a control to assess the model's baseline ability to generate quests without specific guidance, relying solely on the user input prompt.
2. System Prompt 1 (General Guidance): This prompt offers a general guideline for LLM that emphasizes key aspects of the Assassin's Creed Odyssey, including its historical setting, gameplay elements, and narrative structure. It directs the model to generate quests that are contextually appropriate and thematically consistent with the game world.
3. System Prompt 2 (Detailed Instructions): This prompt is more detailed compared to the previous one, that provides detailed instructions for different aspects of the game such as historical accuracy, moral dilemmas, strategic gameplay, complex character dynamics, and immersive environments.
4. System Prompt 3 (Expert Game Designer Persona): This prompt is most detailed instruction that makes LLM adapt the expert game designer role to provide a detailed template for quest design, emphasizing design principles, expected outputs, and creative flexibility within the constraints of historical accuracy and narrative coherence.

3.9.3 Model variations. 12 different models were used in the quest generation process, including both base models (GPT-4o Mini and ChatGPT 3.5 Turbo) and their fine-tuned versions. The fine-tuning process for these models has been designed iteratively to enhance their ability to generate quests that is contextually appropriate and coherent based on the structured data and synthetic datasets developed earlier in the study.

The combination of user input prompts, system role prompts, and model variations resulted in a diverse set of 720 unique quests. This extensive dataset allowed for a comprehensive analysis of how different models and prompt structures influenced the quality and creativity of the generated quests.

3.10 Evaluation Metrics

The evaluation of the generated quests was conducted using a comprehensive set of criteria designed to assess the quality and effectiveness of content produced by fine-tuned Large Language Models (LLMs). The evaluation process involved sending the outputs of different fine-tuned models through a consistent evaluation prompt to various LLMs, specifically Gemini 1.5 Pro, GPT-4o, and GPT-4o Mini, to rate the quest designs. Each criterion was carefully chosen to cover different aspects of game design, narrative coherence, and player engagement, ensuring a thorough assessment of each quest's strengths and weaknesses.

3.10.1 Evaluation prompt. The evaluation was conducted using a crafted prompt designed to simulate the perspective of a game design expert. This prompt directed the LLMs to rigorously assess each quest based on a set of predefined criteria, considering the unique challenges and opportunities presented by LLM-generated content. The evaluation prompt emphasized critical aspects such as narrative coherence, thematic depth, originality, and integration of gameplay mechanics, among others. By applying this prompt consistently across different models, the study aimed to ensure a standardized evaluation process.

3.10.2 Evaluation criteria. The evaluation of quests is based on several criteria. Originality (1-10) assesses the novelty of the quest. A high score indicates that the quest introduces new concepts, settings, or characters that challenge conventional ideas within the game world, offering a fresh perspective that enhances the player's experience. A low score suggests reliance on familiar tropes and clichés, with little innovation in quest design.

Coherence (1-10): Coherence evaluates the logical flow and consistency of the quest's narrative. High coherence is achieved when the conflict, setting, and characters are well-integrated, creating a seamless and immersive story. Low coherence is marked by disjointed narratives, plot holes, or abrupt transitions that disrupt the player's engagement.

Theme Exploration (1-10): This criterion measures the depth with which the quest engages with its central themes. A high score reflects a profound exploration of relevant themes, using dialogue, plot, and gameplay mechanics to challenge the player's thinking. A low score indicates superficial engagement with themes, lacking depth or relevance.

Clarity (1-10): Clarity focuses on how well the quest communicates its objectives to the player. High clarity ensures that the player's goals are clear, intuitive, and aligned with the protagonist's motivations. A low score suggests potential for confusion, with unclear objectives that may mislead the player.

Motivation (1-10): This criterion evaluates how compelling the quest's objectives are for the player. High motivation is achieved when the objectives are meaningful, resonating with the player's interests and driving engagement. Low motivation occurs when the objectives fail to inspire the player to proceed.

Narrative Integration (1-10): Narrative integration assesses how well the quest's objectives contribute to the overall story. A high score indicates that the objectives are central to the narrative, deepening the player's understanding of the story and its

characters. A low score suggests that the objectives feel tangential or disconnected from the main narrative.

Variety (1-10): Variety measures the diversity of gameplay mechanics within the quest. A high score reflects a well-balanced mix of combat, exploration, puzzles, and dialogue, creating an engaging and dynamic experience. A low score indicates monotonous gameplay, relying on repetitive mechanics with little variation.

Integration of Gameplay Mechanics (1-10): This criterion evaluates how well the gameplay mechanics enhance the quest's narrative and themes. High integration occurs when the mechanics feel natural and cohesive within the quest, contributing to the overall experience. Poor integration is marked by mechanics that disrupt the flow of the narrative or detract from the quest.

Challenge (1-10): Challenge assesses the balance and engagement level of the quest's gameplay. A high score indicates appropriately challenging gameplay that is neither too easy nor excessively frustrating, enhancing the player's experience. A low score reflects a lack of balance, resulting in a gameplay experience that may be unsatisfying.

Immersion (1-10): Immersion evaluates the effectiveness of the quest environments in drawing the player into the game world. A high score is given when the settings are detailed, atmospheric, and consistent with the game's tone and setting, fully engaging the player. A low score indicates environments that feel disconnected or poorly designed, breaking the player's immersion.

World-building (1-10): This criterion assesses how well the quest environments contribute to the overall world-building of the game. High world-building adds to the game's lore, history, and atmosphere, making the game world more believable and engaging. A low score suggests environments that feel generic or uninspired, with little contribution to the game's depth.

Dynamic Elements (0-1): Dynamic elements measure the presence of interactive environments that change based on player choices. A score of 1 indicates that the quest includes significant dynamic elements, enhancing replayability and the player's experience. A score of 0 suggests static environments with no significant interactive elements.

Depth (1-10): Depth evaluates the complexity and believability of character interactions within the quest. A high score reflects well-developed characters and meaningful dialogue that contribute to the narrative. A low score indicates shallow or unconvincing interactions.

Historical/Cultural Alignment (1-10): Historical/cultural alignment assesses how accurately the quest reflects the historical period and cultural context of the game. A high score indicates that the quest captures the nuances of the era, including language, customs, and societal norms. A low score reflects anachronistic or culturally inaccurate elements that break immersion.

Relevance (1-10): Relevance evaluates how well the quest's challenges and objectives align with its themes and narrative. A high score is given when the challenges are integral to the story, enhancing the player's understanding of the game's world. A low score indicates challenges that feel arbitrary or disconnected from the larger narrative.

Choice Significance (1-10): Choice significance measures how meaningful the player's choices are in terms of lasting consequences on the story or gameplay. High significance is marked by choices that alter the quest's outcome or impact the broader game world. Low significance indicates choices that have little to no effect on the game's progression.

Replayability (1-10): Replayability assesses the potential for players to revisit the quest, considering elements such as multiple outcomes, branching storylines, and hidden content. A high score reflects strong incentives for replaying the quest, while a low score suggests a linear experience with no incentive to replay.

Resolution (1-10): Resolution evaluates how satisfying the quest's conclusion is. A high score indicates that the resolution feels earned, consistent with the narrative, and provides closure. A low score reflects an unsatisfying resolution, leaving loose ends or a lack of closure.

Lasting Impact (1-10): Lasting impact assesses the long-term effects of the quest within the game world, such as changes to the environment, character relationships, or future quests. A high score indicates a significant impact, leaving a memorable mark on the player's experience. A low score suggests little to no lasting impact.

3.10.3 Structured outputs for documentation. The use of structured outputs was critical in ensuring the consistency and ease of documentation for each evaluation score. The structured approach allowed the LLMs to generate well-organized responses, which facilitated accurate recording and analysis of the evaluation data. This method also made it straightforward to track and compare the performance of different models across the various evaluation criteria.

3.10.4 Bias mitigation through multi-model evaluation. To reduce potential bias in the evaluation process, the same quests generated by different fine-tuned models were evaluated by multiple LLMs, including Gemini 1.5 Pro, GPT-4o. By using different models to assess the same content, the study aimed to ensure that the evaluations were not overly influenced by the characteristics or limitations of any single model. This multi-model approach provided a more balanced and comprehensive assessment of each quest's quality, as it allowed for cross-comparison of scores and highlighted any discrepancies that might arise from model-specific biases.

3.10.5 Penalization guidelines. The evaluation process included specific guidelines for penalizing quests that lacked critical or optional elements. Critical elements, such as narrative coherence and clear objectives, were heavily penalized, with deductions ranging from 2 to 5 points depending on the severity of the omission. Optional elements, such as dynamic environments, were penalized minimally, with a standard 0 or 1 point deduction based on their presence or absence.

3.10.6 Comparative analysis and review. The LLMs were instructed to compare each quest to others within the same game, applying a critical lens to identify standout features and areas for improvement. After assigning initial scores, the LLMs revisited their evaluations to ensure accuracy, reflecting a nuanced and balanced assessment of the quest's overall quality.

3.10.7 Integration into the dataset. The scores generated by the LLMs were systematically recorded and integrated into the dataset. This structured approach enabled a quantitative analysis of the performance of different fine-tuned models, providing insights into the strengths and limitations of LLMs in generating high-quality game quests.

Chapter 4

Findings

The objective of this study was to evaluate the effectiveness and quality of quests generated by different Large Language Models (LLMs) under various conditions. By analysing these conditions and their combinations, the research aims to identify key parameters for effectively using LLMs in video game development. The study utilized 12 different models (10 fine-tuned models and 2 base models) to generate quests from 15 distinct user inputs. Additionally, 4 different system prompts were employed, each providing varying levels of guidance to the models before processing the user inputs. These system prompts progressively increased in detail to guide the models toward producing more refined and contextually appropriate quest designs. In total, 720 different LLM-generated quest outputs were documented and subsequently evaluated according to 19 predefined criteria for assessing AI-generated quest content.

4.1 Evaluation Model Comparison (GPT4o & Gemini 1.5 Pro)

All models contain some sort of bias since each model has been trained by human written text. Also, those models are continued to be trained based on the conversations that have been made regularly. Before delving into key findings, it is important to understand the models' disposition of evaluation criteria.

4.1.1 Mean value comparison across criteria. The mean values show that GPT-4o generally assigned higher scores across most criteria compared to Gemini. GPT-4o tends to be more lenient in its evaluations, giving higher average scores areas like Relevance, Narrative Integration, and Immersion. This could imply that GPT-4o is more favourable in its assessments or perhaps less critical in identifying flaws in these aspects. Conversely, Gemini tends to assign lower mean scores, particularly in Challenge, Variety, and Lasting Impact. This may indicate that Gemini is more stringent in its evaluations, potentially identifying issues that GPT-4o overlooks or considers less severe.

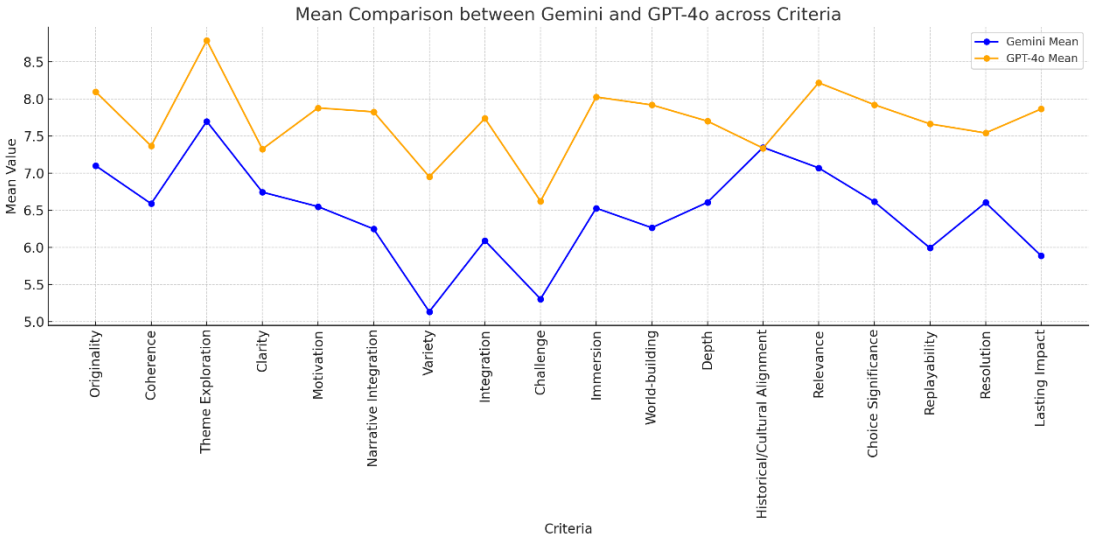

Figure 3 Mean scores of quest evaluations by Gemini and GPT-4o across different criteria.

4.1.2 Standard deviation comparison across criteria. The standard deviation comparison between GPT-4o and Gemini 1.5 Pro provides insight into the consistency of the models' evaluations across different criteria. As illustrated in the graph, there are noticeable differences in how each model scores the quests, reflecting variability in their evaluation processes.

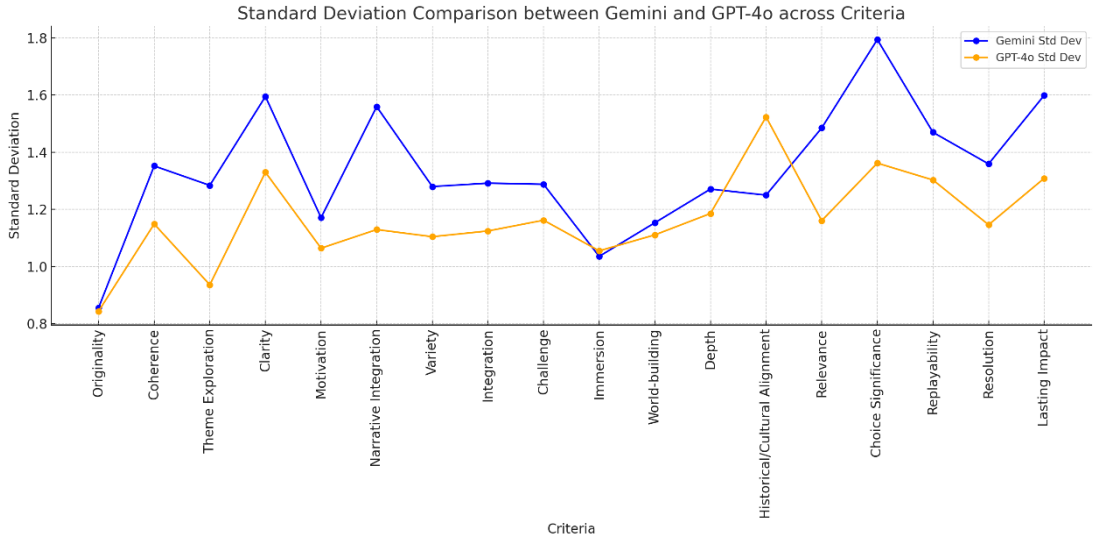

Figure 4 Standard deviation comparison of evaluated quests by Gemini and GPT-4o across different criteria.

Gemini generally exhibits a higher standard deviation across most criteria compared to GPT-4o. This suggests that Gemini's scoring is more variable, indicating scoring range is wider compared to GPT-4o evaluations. For instance, criteria such as Clarity, Narrative Integration, and Choice Significance show significant variability in Gemini's scores, with standard deviations exceeding 1.4 in some cases. This high variability might reflect a more nuanced or fluctuating evaluation approach, where Gemini's assessments are more sensitive to differences in the quest content.

On the other hand, GPT-4o displays a more stable evaluation pattern, with lower standard deviations in most criteria. This suggests that GPT-4o tends to assign scores more consistently across different quests, particularly in the upper scoring side. This potentially results in less differentiation between the quality of individual quest elements.

4.1.3 Frequency distribution of criteria scores. The frequency distributions highlight distinct differences in how GPT-4o and Gemini evaluate the same quests. GPT-4o generally assigns higher scores across most criteria, indicating a more lenient or favorable assessment. In contrast, Gemini's scores are more variable and often concentrated in the mid-range, reflecting a more critical and nuanced evaluation approach as discussed in previous sections. This tendency of GPT-4o to award higher scores may be influenced by its closer alignment with the base and fine-tuned GPT models that generated the quest outputs.

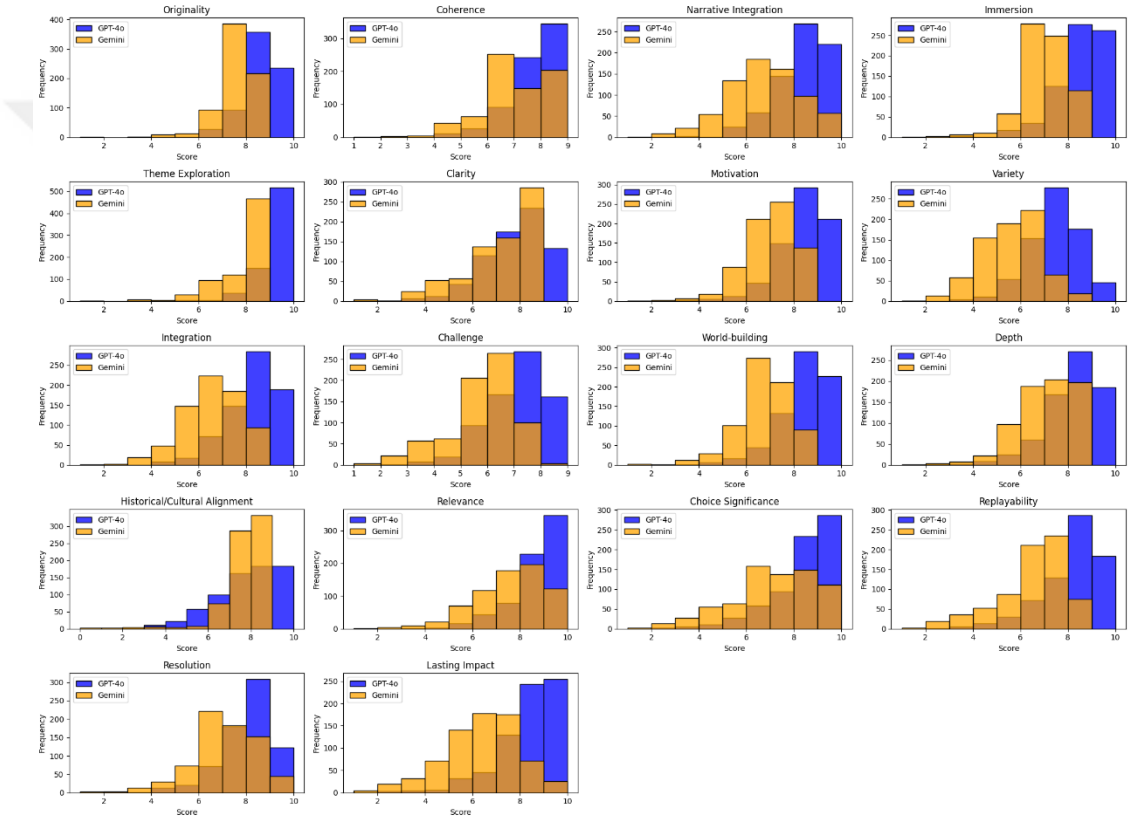

Figure 5 Frequency distribution of scores across criteria by GPT-4o and Gemini.

4.1.4 Kurtosis comparison and weighting. Kurtosis is a statistical measure that helps us understand the "tailedness" or extremity of a distribution (DeCarlo, 1997). In simpler terms, kurtosis indicates how much of the data is in the tails of the distribution versus around the mean. High kurtosis suggests that the data has heavy tails, meaning more of the scores are found at the extremes (very high or very low), indicating less consistency. Low kurtosis, on the other hand, indicates that the scores are more clustered around the mean, suggesting more consistency and less variability in the scoring.

In this study, kurtosis was employed as a key metric to assess the consistency of scores given by two different Large Language Models (LLMs), GPT-4o and Gemini 1.5 Pro, across various evaluation criteria. Each criterion was scored by both models on a scale of 1 to 10. However, instead of merely comparing the mean scores to draw conclusions about the models' performances, we recognized the importance of considering the distribution of these scores—how consistently each model applied its scoring across different quest outputs.

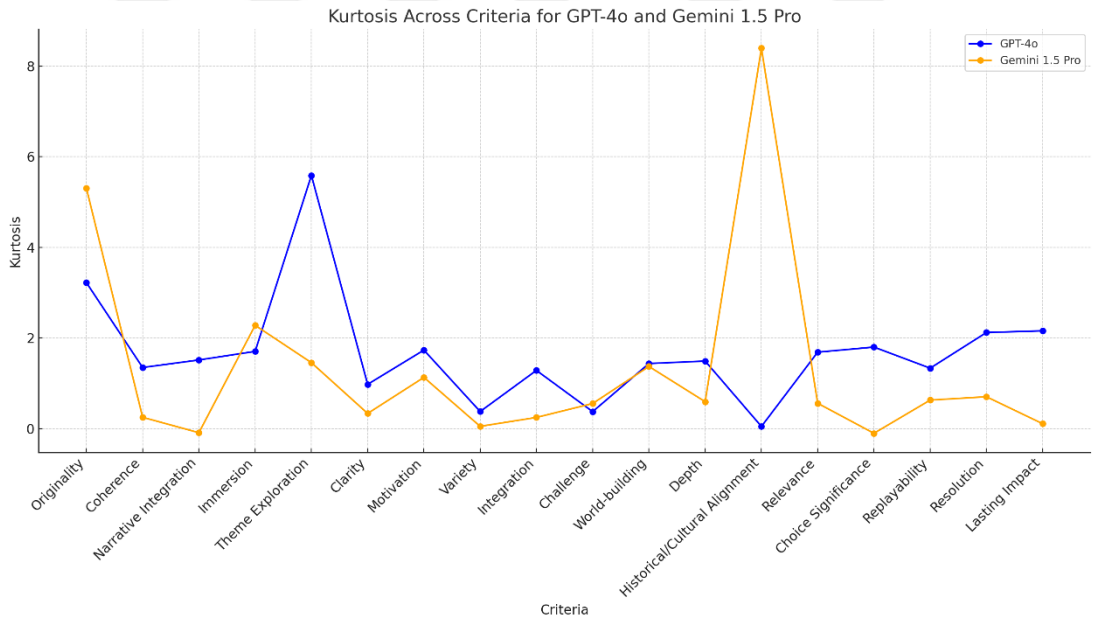

Figure 6 Kurtosis comparison across criteria for GPT-4o and Gemini 1.5 Pro.

The kurtosis values were used to determine weights for combining scores from the two models. Specifically, the weights were based on the inverse of the kurtosis values, such that models exhibiting greater consistency (lower kurtosis) were assigned

higher weights. This approach ensured that the final weighted scores reflected more reliable and stable evaluations, minimizing the influence of models with less consistent scoring patterns.

Additionally, one of the criteria in this study was derived from the evaluations of both LLMs, which further underscores the importance of weighting these scores appropriately. By considering the consistency of each model's scoring through kurtosis, the study avoids the pitfalls of simple averaging, which could otherwise misrepresent the true performance of the models. This method ensures that the final weighted scores not only reflect the average quality as perceived by the models but also incorporate the degree of confidence in each model's evaluations. This methodology could be further extended to other contexts where multiple evaluative models are used, ensuring that the most consistent and reliable assessments are given greater importance in the analysis.

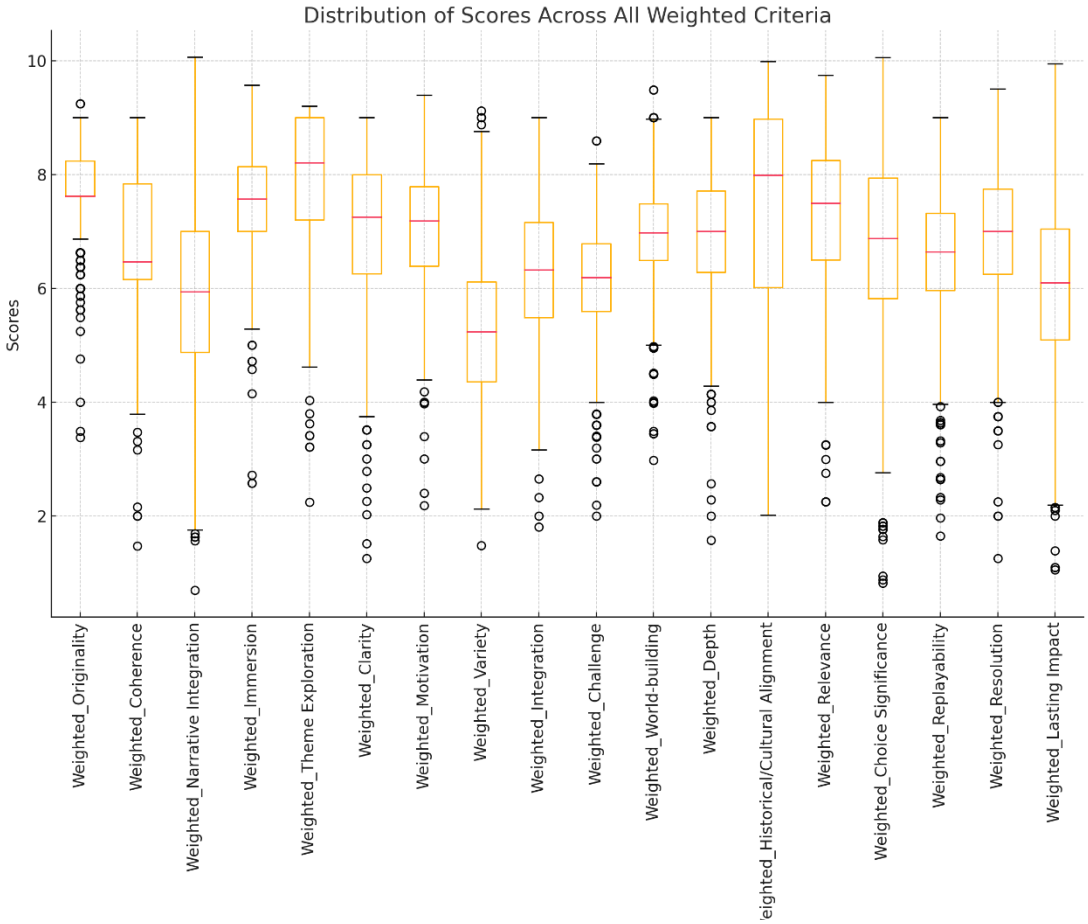

Figure 7 Distribution of scores across all weighted criteria, derived using kurtosis-based weighting.

4.2 Model Performance Comparison

By weighting the criteria, we arrive at a single, adjusted score for each criterion, simplifying the performance analysis. While the idea of comparing different evaluation models for each category is intriguing, for the scope of this research, the focus will be on the combined, weighted scores to provide a more holistic assessment of the models' performances.

4.2.1 Base models vs fine-tuned models. To compare the performance of the models, we introduced a feature representing the total score of the quests generated by the Large Language Models (LLMs). This total score is derived by summing up the scores across all evaluation criteria, providing a generic measure of the model's output quality. This aggregated score allows for a comprehensive comparison between the base models and their fine-tuned counterparts, enabling us to identify which models consistently produce higher-quality quests.

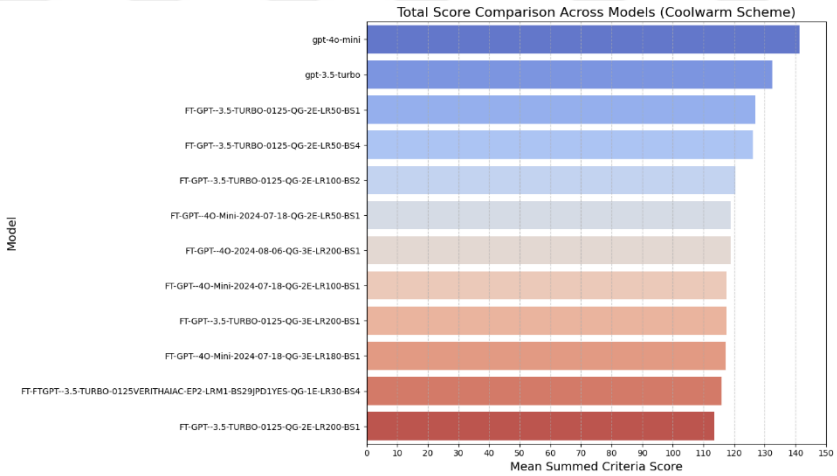

Figure 8 Total score comparison between base and fine-tuned models.

Contrary to typical expectations, it is evident that base models perform consistently better compared to any other fine-tuned models. The fine-tuned models exhibit a range of performance scores, generally lower than the base models. The top fine-tuned model is "FT-GPT--3.5-TURBO-0125-QG-2E-LR50-BS1," which still

scores lower than the base models, implying that fine-tuning with these specific configurations did not enhance the performance as expected and, in some cases, may have slightly degraded the output quality.

This observation is further supported by the distribution of total scores across models, as illustrated in the box plot. The base models, particularly "gpt-3.5-turbo" and "gpt-4.0-mini," not only have higher median scores but also demonstrate a more consistent performance with relatively narrower interquartile ranges. In contrast, the fine-tuned models show greater variability in their score distributions, with several models exhibiting wider interquartile ranges and the presence of outliers. These outliers indicate instances where fine-tuned models significantly underperformed, reinforcing the notion that fine-tuning, under certain configurations, introduced inconsistencies in model output quality. The broad spread of scores among fine-tuned models suggests that while fine-tuning can adjust model performance, it may also introduce unpredictability, underscoring the complexity and challenges associated with optimizing LLMs for specific tasks like quest generation.

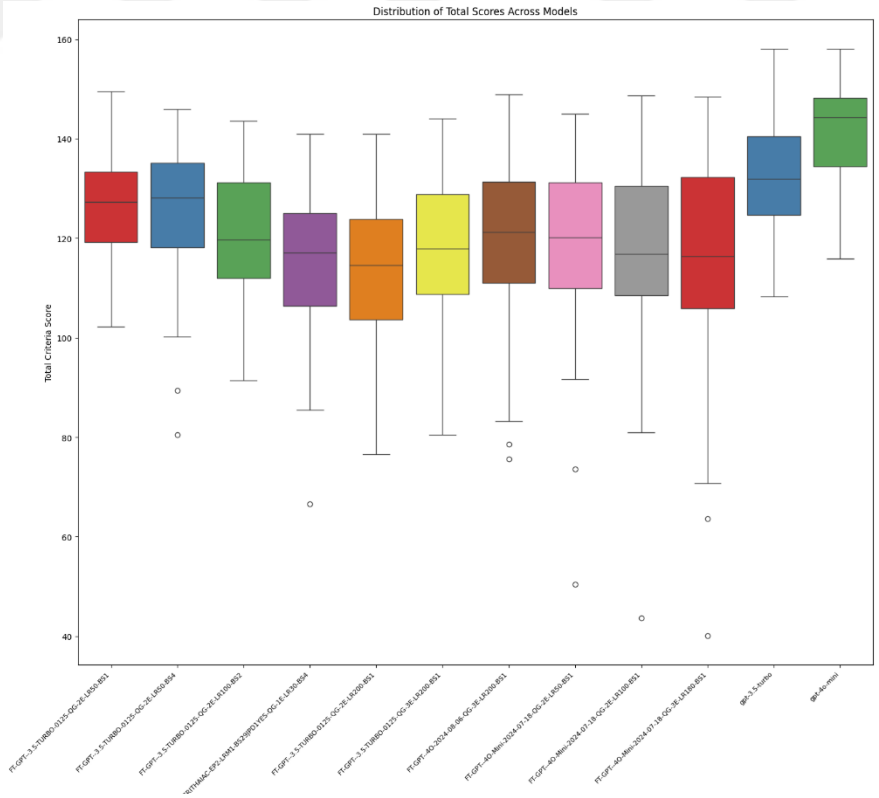

Figure 9 Box plot showing the distribution of total scores across fine-tuned and base models.

For all evaluation criteria, the fine-tuned models underperform compared to base models, as indicated by the negative differences shown in chart. The differences, while not massive in all cases, are statistically significant and consistently favor the base models.

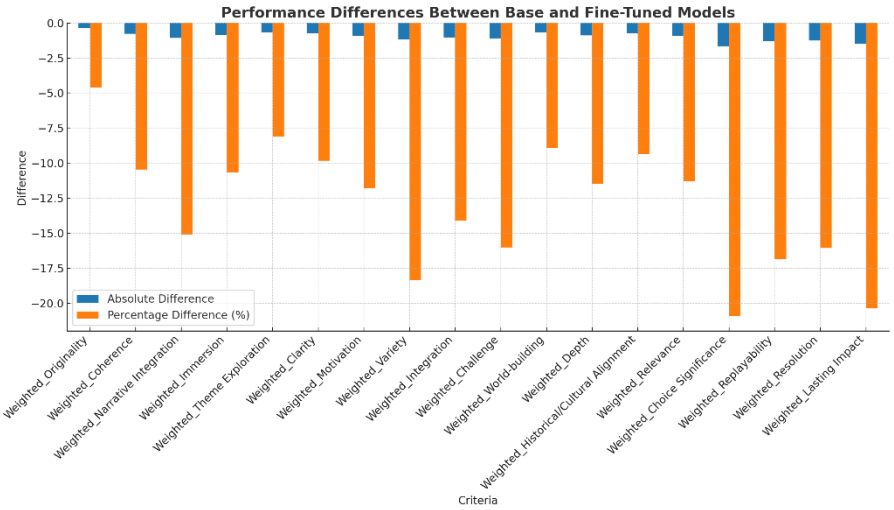

Figure 10 Bar chart showing the performance differences between base and fine-tuned models across various criteria, illustrating both absolute and percentage differences.

Weighted Choice Significance shows the most substantial decline in performance for fine-tuned models, with a percentage difference of over -20%. Weighted Originality and Weighted Theme Exploration have relatively small differences, suggesting that fine-tuned models manage to retain some performance in these areas. The consistent underperformance of fine-tuned models suggests that the current fine-tuning strategy may need to be revised.

4.2.2 Performance within fine-tuned models. In examining the performance within fine-tuned models, it becomes evident that the variations in hyperparameters such as batch size, learning rate, and the number of training epochs significantly influence the outcomes across the 18 evaluation criteria. One of the most consistent findings is related to the criterion of Originality, where the fine-tuned models display relatively stable performance. Despite slight variations, the low standard deviations suggest that different fine-tuning approaches, particularly variations in batch sizes, lead to a uniform level of originality in the generated quests.

Batch size, while influential, does not appear to be the most critical factor in determining output quality. Smaller batch sizes tend to perform slightly better in criteria like Originality and Immersion, yet these differences are modest. This indicates that while batch size contributes to overall performance, its impact is somewhat overshadowed by other hyperparameters, such as the learning rate.

The learning rate, on the other hand, plays a pivotal role in shaping the model's effectiveness, particularly in coherence and narrative integration. Models fine-tuned with a lower learning rate, such as 0.5, generally perform better across these criteria. This suggests that a more cautious approach to learning allows the model to avoid overfitting, enabling it to generalize better to new data. Conversely, higher learning rates, although they facilitate faster convergence, often result in less refined outputs, as evidenced by a slight drop in mean scores. The variability in performance, as indicated by standard deviation, highlights that the learning rate is a critical determinant in achieving consistent and reliable results. Also, higher learning rates show a decline in performance, underscoring the importance of careful tuning.

The number of training epochs is another significant factor that influences model performance. Models subjected to more extended training generally perform better in areas like Coherence and Clarity. This improvement suggests that prolonged training periods allow the model to fine-tune its outputs more effectively. However, the gains from additional epochs tend to diminish beyond a certain point, indicating that while more epochs contribute to consistency, there is a threshold after which the benefits plateau.

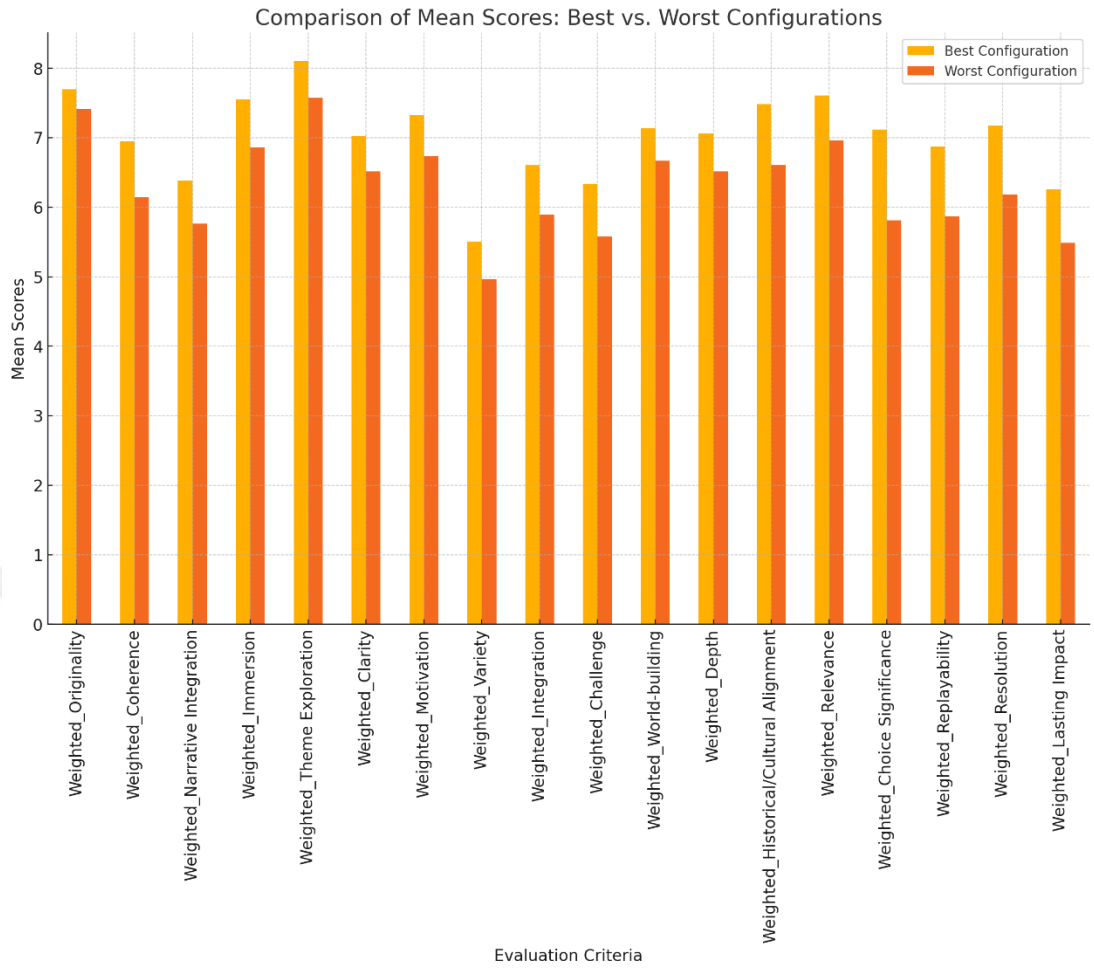

Figure 11 Bar chart showing comparison of mean scores across various evaluation criteria between best and worst fine-tuning configurations.

When comparing the best and worst-performing configurations, it becomes clear that hyperparameter optimization is crucial. The best configuration, characterized by a well-balanced learning rate and an optimal number of epochs, resulted in consistently higher mean scores, especially in Immersion, Theme Exploration, and Narrative Integration. This contrasts sharply with the worst configuration, where higher learning rates or insufficient training epochs led to significantly lower scores in Challenge and Clarity. These findings underscore the importance of carefully tuning hyperparameters to achieve high-quality outputs, particularly in criteria that are sensitive to the nuances of model training.

It is important to note that these findings are reliant on the synthetic dataset used for fine-tuning models. While some observations may have general applicability,

others may be specifically constrained by the dataset's scope, and thus, broader generalization should be approached with caution.

4.2.3 System prompt impact on model performance. In this section, we will analyze the impact of different system prompts on the performance of the models. System prompts are the instructions or guidelines provided to the models to guide their behavior when generating responses. These prompts can significantly influence the quality and consistency of the model outputs.

There are 4 different system prompts that have been used. At the most basic level (no system message), the model operates without any guiding instructions, relying solely on the input prompt provided by the user. A slightly more detailed system prompt (System Message Category 1) might include broad guidelines that align the quest with the game's themes, player motivations, and environment, ensuring a basic level of coherence and relevance to the game's world. As the level of detail increases (System Message Category 2), the prompt could include specific instructions for integrating historical conflicts, moral dilemmas, and dynamic environments into the quest, providing a richer and more immersive experience. The most detailed system prompt (system message 3) would provide comprehensive instructions that guide the model to design quests with deep narrative complexity, historical accuracy, and strategic gameplay elements, ensuring that every aspect of the quest is meticulously crafted to enhance both the storytelling and player engagement.

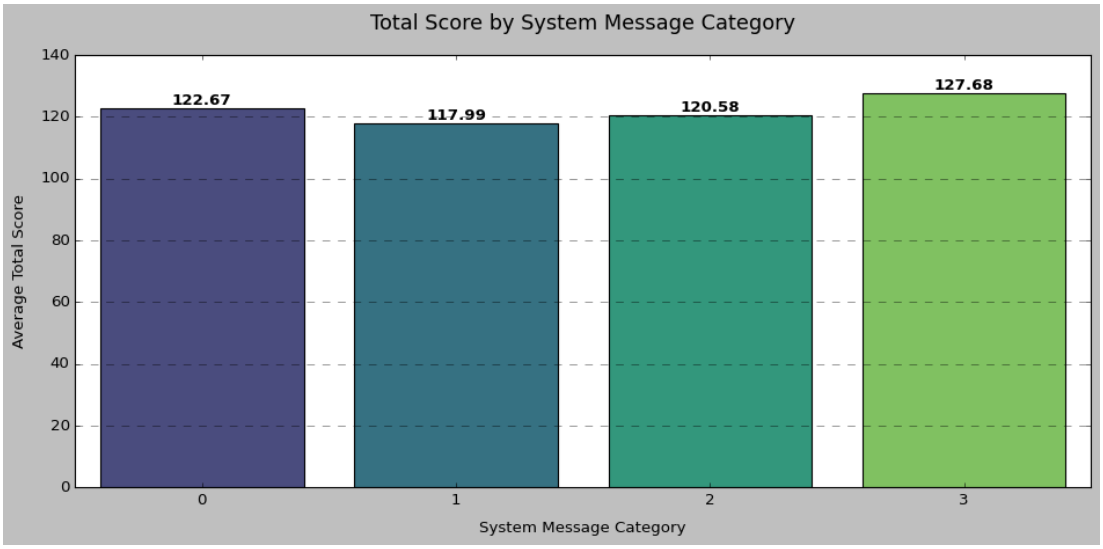

Figure 12 Bar chart illustrating the average total scores across different system message categories.

While the most detailed system prompt generally yields better total scores, it's crucial to examine how different models respond to these prompts.

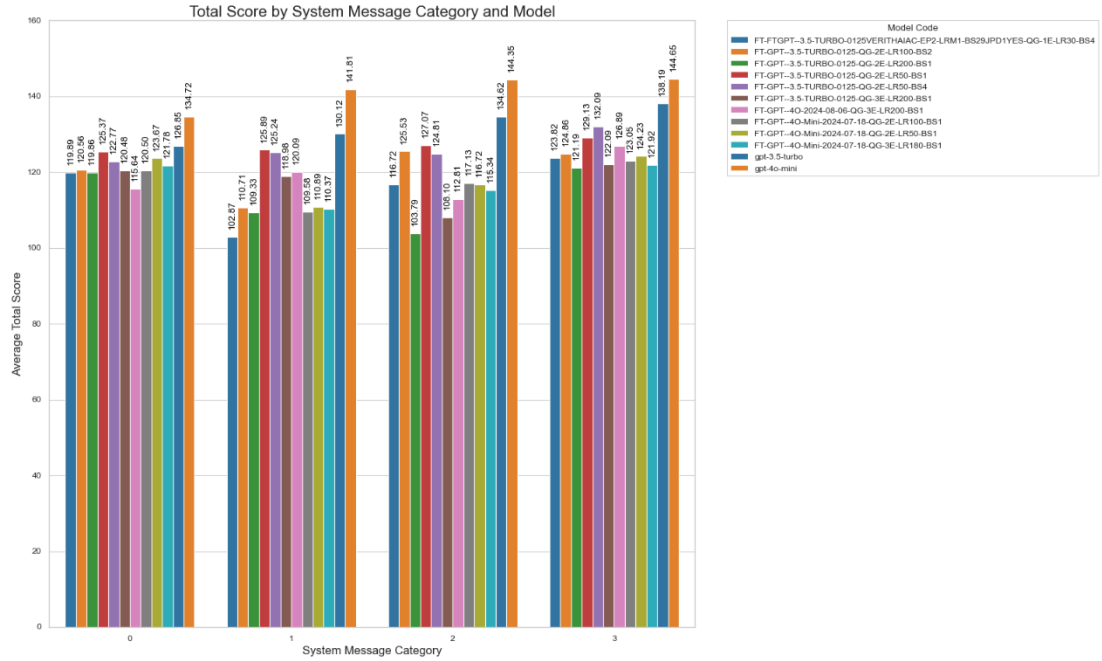

Figure 13 Bar chart displaying the average total scores for each model across different system message categories.

Base models, which operate without fine-tuning, generally show better performance across various metrics. However, the performance gap between base and

fine-tuned models diminishes when no system prompt is used. This suggests that base models may derive more significant benefits from system prompts compared to fine-tuned models. In other words, base models appear to be more sensitive to the guidance provided by system prompts, leveraging these instructions to enhance their output quality more effectively than fine-tuned models.

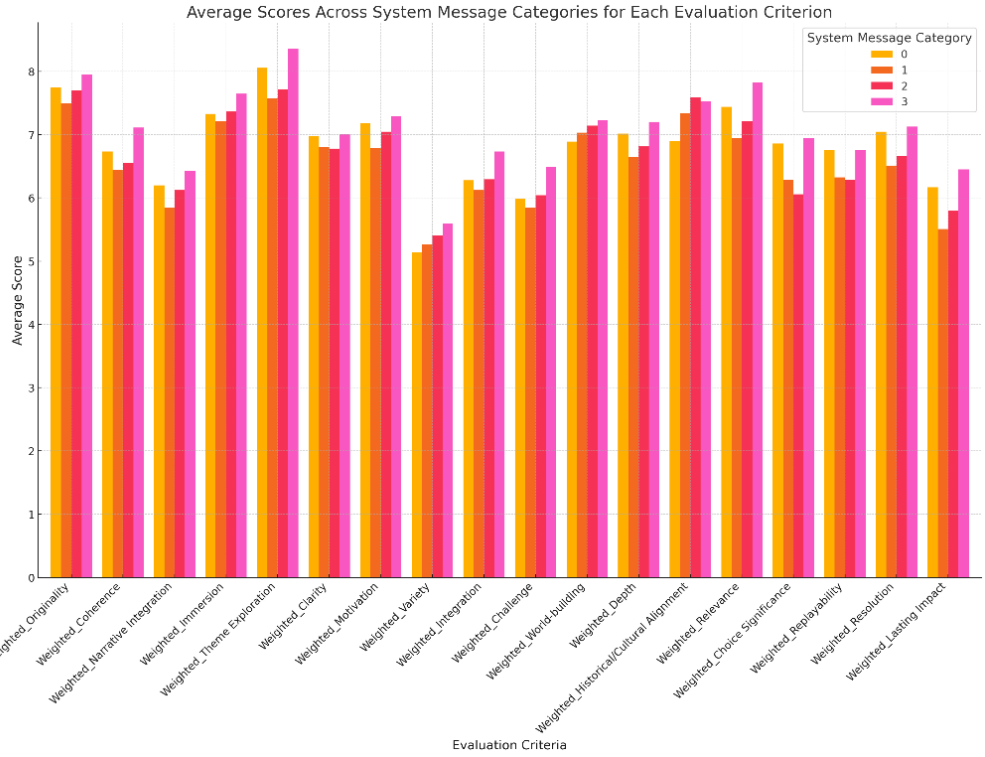

Figure 14 Average scores across system message categories for each evaluation criterion.

While detailed system prompts (e.g., System Message Category 3) tend to improve model performance, this improvement is not consistently proportional to the level of detail. Interestingly, in some criteria, models produced better outputs without any system prompt compared to when using System Message Categories 1 and 2. This finding suggests that the effectiveness of prompts may not always increase with added complexity. Instead, the results imply that the benefits of system prompts might follow an all-or-none pattern, similar to the activation function of McCulloch–Pitts neurons discussed earlier in this thesis. That is, unless a certain threshold of prompt detail is met, additional complexity may not provide substantial performance gains.

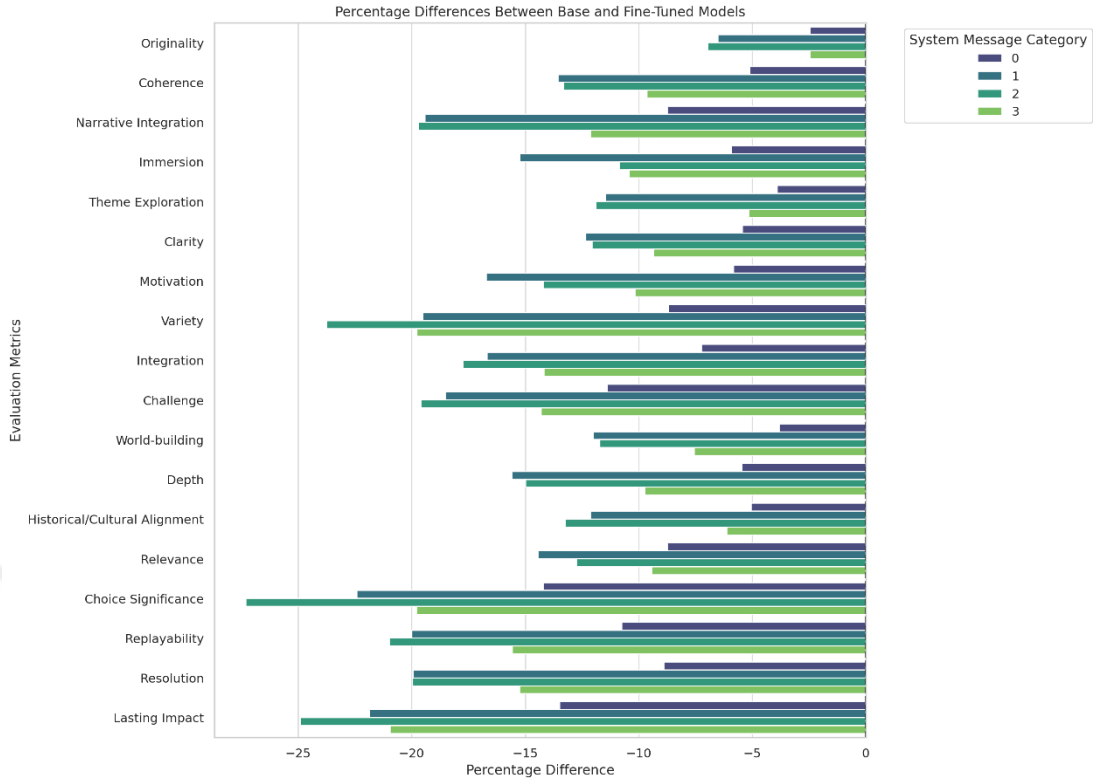

Figure 15 Percentage differences between base and fine-tuned models across evaluation metrics, categorized by system message category.

The accompanying chart visually represents the percentage differences in performance between base and fine-tuned models across various evaluation metrics, categorized by the four different system prompts. It highlights that, while base models generally outperform fine-tuned models, the degree of this advantage varies significantly depending on the system prompt used. Notably, the chart reveals that as the detail in system prompts increases, the performance gap between base and fine-tuned models often widens in favor of base models. This trend suggests that base models benefit more from the additional guidance provided by detailed system prompts. However, it is important to note that this benefit is not uniform across all metrics. This further emphasizes the non-linear relationship between prompt complexity and model performance, underscoring that the effectiveness of system prompts is not solely dependent on the level of detail, but rather on how well the prompts align with the model's capabilities and the specific criteria being evaluated.

Chapter 5

Discussions and Conclusions

We are on the verge of a significant transformative era shaped by artificial intelligence. These advancements are progressing rapidly, often building upon existing knowledge in ways that can quickly render previous research obsolete. This acceleration highlights the importance of bridging the gap between academic insights and real-time applications, especially for those outside the academic community. Ensuring that AI developments are accessible and understandable is crucial for maintaining social justice in a rapidly changing world.

Given this context, the present study seeks to contribute to this evolving landscape by exploring potential approaches. The number of research papers related to procedural quest generation and LLMs will continue to increase, and this research paper aims to serve as a guiding light. The goal was not merely to find answers but also to create new questions that have not been asked before. The evolving landscape of AI allows us to explore different methodologies from various disciplines, and the future awaits exploration at the intersection of these domains.

These rapid developments will increase the demand for data from any available sources. Most high-quality data have been provided at a cost, and the accessibility of such data will become even more important as demand increases. This highlights the significance of sources like Fandom Wiki, which provides data collected through collective labor. The amount of collective labor can be seen as “spreading knowledge” to all people, which is an important aspect of being civil and feeling civil. These aspects are vital for democratizing knowledge, which may become a challenge in the future as private LLM companies gain the power to limit, direct, or alter knowledge at will—making it difficult to audit these processes. In addition to expressing gratitude for the collective labor—specifically the dataset from the Assassin’s Creed Odyssey Wiki used in this research—there is always room for improvement in how we collect data and how we relate this data to create value.

This research is highly dependent on the data taken from Fandom Wiki, which had to be structured for further analysis. Since converting data into a structured form is a time-consuming process, the tools that have been designed will be shared for "spreading knowledge" purposes. These tools will create ease in parsing the data from Fandom Wiki with small adjustments. Even though these processes may seem irrelevant to fine-tuning, the most important aspect of fine-tuning is how we cluster the information and how we simulate a conversation that adjusts the model's weights and biases. Without the ability to structure data, it is nearly impossible to conduct research related to fine-tuning.

The findings from this study reveal that fine-tuned models often exhibit noticeable degradation in performance compared to base models, with even the best fine-tuned model failing to match the output quality of GPT-3.5 Turbo. This degradation is largely attributed to the limitations inherent in the fine-tuning dataset, which was synthetic and generated using different LLMs. The output, simulating the assistant reply, was shaped by Alpay's (2022) framework, which categorized tasks into five distinct categories: termination, retrieval, delivery, communication, and arrival.

While this framework provided a structured approach to quest design, it inadvertently constrained the models' ability to make associations with other concepts, leading to a more static and predictable output. As a result, fine-tuned models struggled with criteria such as Choice Significance and Lasting Impact, where creativity and dynamic narrative elements are crucial. This static nature, imposed by the rigid task clustering within the quest framework, hindered the models' ability to fully leverage the potential of fine-tuning.

To overcome these limitations and unlock the true potential of fine-tuning, there is a compelling need for a more advanced approach. Instead of fitting all quest designs into a single, static structure, a pipeline approach should be employed. This pipeline would involve connecting multiple fine-tuned models, each specialized in a narrow aspect of quest design. By allowing these models to transform and exchange

information, the pipeline could integrate various perspectives, enhancing the overall creativity and complexity of the narratives produced.

Such a system would not only mitigate the predictability introduced by fine-tuning but also foster a more dynamic and flexible quest generation process. The interdisciplinary nature of video games and AI, while labor-intensive, offers significant potential when approached through collaborative efforts. By leveraging the expertise of communities, effectively administered teams, or interdisciplinary groups of academicians, the video game industry can develop sophisticated model pipelines that push the boundaries of procedural content generation.

Some may argue that the evaluation may contain biases since the research was conducted mostly using LLMs. Interestingly, each LLM model has been trained on human-written text, which itself contains biases, and it could also be argued that humans tend to have biases as well. This raises the question: does this lead us to critique human evaluation as well? While human evaluation and expert reviews were conducted in an unorganized manner in this research, the power of large sample sizes is undeniable. For this research, 720 different quests were generated by different model combinations and evaluated twice, once by GPT-4o and once by Gemini 1.5 Pro. Human evaluation could only be conducted for a smaller number of quests. LLMs' tendency to give different results, even when all combinations are the same, is real. Therefore, the validity of the evaluation was increased by the sample size. Evaluating 720 quests on 19 different criteria is a time-consuming task, but using LLMs made the evaluation process cost $25, which is equivalent to 1-2 hours of work in developed countries. Evaluation for different domains with well-defined criteria is an unexplored territory that needs to be discovered.

Future research should consider integrating a broader array of data sources beyond Fandom Wiki to enhance the diversity and comprehensiveness of training datasets. Also, there is a critical need for continued interdisciplinary collaboration to push boundaries of what LLM’s can achieve in creative industries. Partnerships between computer scientists, narrative designers, and game developers could lead to more innovative approaches and solutions.

This research underscores the vast potential of AI in transforming creative processes within the video game industry. However, it also highlights the challenges- such as data dependency and the need for precise model tuning- that must be addressed to fully harness this potential. It is hoped that the methodologies and insights derived from this study will inspire further research and practical applications.

REFERENCES

Abraham, T. H. (2002). (Physio)logical circuits: The intellectual origins of the McCulloch–Pitts neural networks. Journal of the History of the Behavioral Sciences, 38(1), 3–25. https://doi.org/10.1002/jhbs.1094

Abrahamson, C.E. (1998). Storytelling as a Pedagogical Tool in Higher Education. Education 3-13, 118, 440.

Al-Nassar, S., Schaap, A., Zwart, M. V. D., Preuss, M., & Gómez-Maureira, M. A. (2023). QuestVille: Procedural Quest Generation Using NLP Models. Proceedings of the 18th International Conference on the Foundations of Digital Games. Presented at the Lisbon, Portugal. doi:10.1145/3582437.3587188

Alpay, C. (2022). Correlation between the quest in literature & the quest in single-player CRPGs: A look at quests of The Witcher 3: Wild Hunt [Master's thesis, Bahçeşehir University]. Bahçeşehir University Graduate School.

Ammanabrolu, P., Broniec, W., Mueller, A., Paul, J., & Riedl, M. (. oct # "--3 " # nov 2019). Toward Automated Quest Generation in Text-Adventure Games. In B. Burtenshaw & E. Manjavacas (Eds.), Proceedings of the 4th Workshop on Computational Creativity in Language Generation (pp. 1–12). Retrieved from https://aclanthology.org/2019.ccnlg-1.1

Anderson, J. A. (2006). McCulloch–Pitts neurons. Encyclopedia of Cognitive Science. https://doi.org/10.1002/0470018860.s00353

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., Dassarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T.B., Clark, J., McCandlish, S., Olah, C., & Kaplan, J. (2021). A General Language Assistant as a Laboratory for Alignment. ArXiv, abs/2112.00861.

Banerjee, S., & Lavie, A. (2005, June). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 65–72). Retrieved from https://aclanthology.org/W05-0909

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Carli, D. M. D., Bevilacqua, F., Tadeu Pozzer, C., & d'Ornellas, M. C. (2011). A Survey of Procedural Content Generation Techniques Suitable to Game Development. 2011 Brazilian Symposium on Games and Digital Entertainment, 26–35. doi:10.1109/SBGAMES.2011.15

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15, 1 - 45.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. Neural Information Processing Systems.

Chiang, C., & Lee, H. (2023). Can Large Language Models Be an Alternative to Human Evaluations? Annual Meeting of the Association for Computational Linguistics.

Choi, Y., Kim, S., & Lee, J.-H. (2016). Recurrent Neural Network for Storytelling. 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), 841–845. doi:10.1109/SCIS-ISIS.2016.0182

Chu, H. and Hwang, G. (2008). A delphi-based approach to developing expert systems with the cooperation of multiple experts. Expert Systems With Applications, 34(4), 2826-2840. https://doi.org/10.1016/j.eswa.2007.05.034

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021, August). All That's `Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 7282–7296). doi:10.18653/v1/2021.acl-long.565

Connor, A. M., Greig, T. J., & Kruse, J. (2017). Evaluating the impact of procedurally generated content on game immersion. The Computer Games Journal, 6(4), 209-225. https://doi.org/10.1007/s40869-017-0043-6

Crouch, G. I. (1991). Expert computer systems in tourism: emerging possibilities. Journal of Travel Research, 29(3), 3-10. https://doi.org/10.1177/004728759102900301

DeCarlo, L. T. (1997). On the meaning and use of kurtosis.. Psychological Methods, 2(3), 292-307. https://doi.org/10.1037/1082-989x.2.3.292

Doran, J., & Parberry, I. (2011). A prototype quest generator based on a structural analysis of quests from four MMORPGs. Proceedings of the 2nd International Workshop on Procedural Content Generation in Games. Presented at the Bordeaux, France. doi:10.1145/2000919.2000920

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

Dreyfus, S.E. (1990). Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. Journal of Guidance Control and Dynamics, 13, 926-928.

Emami, A. and Jelinek, F. (2005). A neural syntactic language model. Machine Learning, 60(1-3), 195-227. https://doi.org/10.1007/s10994-005-0916-y

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A Bibliometric Review of Large Language Models Research from 2017 to 2023. ACM Transactions on Intelligent Systems and Technology.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017, July). Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 634–642). doi:10.18653/v1/P17-1059

Hadi, M. U., Tashi, Q. A., Shah, A., Qureshi, R., Muneer, A., Irfan, M., … & Shah, M. (2024). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects.. https://doi.org/10.36227/techrxiv.23589741.v5

Harris, S., & Caldwell, N. (2023). A transfiguration paradigm for quest design. Games and Culture, 155541202311701. https://doi.org/10.1177/15554120231170152

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? Science, 298(5598), 1569-1579. https://doi.org/10.1126/science.298.5598.1569

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. ArXiv, abs/1902.00751.

Howard, J. (2008). Quests: Design, Theory, and History in Games and Narratives (1st ed.). A K Peters/CRC Press. https://doi.org/10.1201/b10929

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

http://dx.doi.org/10.1093/mind/LIX.236.433

https://doi.org/10.4324/9780203328064

Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv, abs/2106.09685.

Huizinga, J. (1950). Homo Ludens: A study of the play-element in culture. Beacon Press.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America, 62(S1), S63-S63.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., & Fung, P. (2022). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55, 1 - 38.

Karpinska, M., Akoury, N., & Iyyer, M. (2021, November). The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 1265–1285). doi:10.18653/v1/2021.emnlp-main.97

Katz, D., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). Gpt-4 passes the bar exam. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4389233

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS digital health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Kybartas, B., & Bidarra, R. (2017). A Survey on Story Generation Techniques for Authoring Computational Narratives. IEEE Transactions on Computational Intelligence and AI in Games, 9(3), 239–253. doi:10.1109/TCIAIG.2016.2546063

Lawrence, S., Giles, C. L., & Fong, S. (2000). Natural language grammatical inference with recurrent neural networks. IEEE Transactions on Knowledge and Data Engineering, 12(1), 126-140. https://doi.org/10.1109/69.842255

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: a convolutional neural-network approach. IEEE Transactions on Neural Networks, 8(1), 98-113. https://doi.org/10.1109/72.554195

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R., Hubbard, W., … & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541

Lima, E.S., Feijó, B., & Furtado, A.L. (2019). Procedural Generation of Quests for Games Using Genetic Algorithms and Automated Planning. 2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames), 144-153.

Lin, C.-Y. (2004, July). ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 74–81. Retrieved from https://aclanthology.org/W04-1013

McCarthy J. (1984). Some expert systems need common sense. Annals of the New York Academy of Sciences, 426, 129–137. https://doi.org/10.1111/j.1749-6632.1984.tb16516.x

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943). https://doi.org/10.1007/BF02478259

Melanie Mitchell; July 13–18, 2020. "Conceptual Abstraction and Analogy in Artificial Intelligence." Proceedings of the ALIFE 2020: The 2020 Conference on Artificial Life. ALIFE 2020: The 2020 Conference on Artificial Life. Online. (pp. pp. 7). ASME. https://doi.org/10.1162/isal_a_00354

Merrick, K. (2008). Modeling motivation for adaptive nonplayer characters in dynamic computer game worlds. Computers in Entertainment, 5(4), 1-32.

Minsky, M., & Papert, S. (1969). Perceptrons. M.I.T. Press.

Mishra, M. K. (2023). Generating video game quests from stories University of Twente

Mühlenbein, H. (2009). Computational Intelligence: The Legacy of Alan Turing and John von Neumann. In: Mumford, C.L., Jain, L.C. (eds) Computational Intelligence. Intelligent Systems Reference Library, vol 1. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01799-5_2

Muriel, D., & Crawford, G. (2018). Video Games as Culture: Considering the Role and Importance of Video Games in Contemporary Society (1st ed.). Routledge. https://doi.org/10.4324/9781315622743

Nagy, G. (1991). Neural networks-then and now. IEEE transactions on neural networks, 2 2, 316-8 .

Nasir, M. U., James, S., & Togelius, J. (2024, May 6). Word2World: Generating Stories and Worlds through Large Language Models. arXiv.org. https://arxiv.org/abs/2405.06686

Nay, J.J. (2023). Large Language Models as Fiduciaries: A Case Study Toward Robustly Communicating With Artificial Intelligence Through Legal Standards. ArXiv, abs/2301.10095.

Novikova, J., Dusek, O., Curry, A.C., & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. Conference on Empirical Methods in Natural Language Processing.

Ong, W. (1982). Orality and Literacy: The Technologizing of the Word. Routledge.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). doi:10.3115/1073083.1073135

Pinker, S. (2014). The language instinct: How the mind creates language (Unabridged ed.). Brilliance Audio.

Pospíchal, J., & Kvasnička, V. (2015). 70th Anniversary of Publication: Warren McCulloch & Walter Pitts - A Logical Calculus of the Ideas Immanent in Nervous Activity. In Advances in intelligent systems and computing (pp. 1–10). https://doi.org/10.1007/978-3-319-10783-7_1

Rao, A.S., Kim, J.N., Kamineni, M., Pang, M., Lie, W., & Succi, M.D. (2023). Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv : the preprint server for health sciences.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408.

Ruder, S. (2019). Neural Transfer Learning for Natural Language Processing. ,

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

Sasaki, Y. (2007). The truth of the f-measure. 2007. URL: https://www. cs. odu. edu/mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf [accessed 2021-05-26], 49.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-424.

Shortliffe EH. Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. Proc Annu Symp Comput Appl Med Care. 1977 Oct 5:66–9. PMCID: PMC2464549.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ArXiv, abs/1906.02243.

Todd, G., Earle, S., Nasir, M. U., Green, M. C., & Togelius, J. (2023). Level Generation Through Large Language Models. Proceedings of the 18th International Conference on the Foundations of Digital Games. Presented at the Lisbon, Portugal. doi:10.1145/3582437.3587211

Togelius, J., Champandard, A. J., Lanzi, P. L., Mateas, M., Paiva, A., Preuss, M., & Stanley, K. O. (01 2013). Procedural content generation: goals, challenges and actionable steps. Dagstuhl Follow-Ups, 6, 61–75.

Turing, A.M. (1950) Computing Machinery and Intelligence. Mind, 59, 433-460.

van Stegeren, J., & Myśliwiec, J. (2021). Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. Proceedings of the 16th International Conference on the

Foundations of Digital Games. Presented at the Montreal, QC, Canada. doi:10.1145/3472538.3472595

Värtinen, S., Hämäläinen, P., & Guckelsberger, C. (2024). Generating Role-Playing Game Quests With GPT Language Models. IEEE Transactions on Games, 16(1), 127–139. doi:10.1109/TG.2022.3228480

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv, abs/2201.11903.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9, 36 - 45. https://doi.org/10.1145/365153.365168

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. Neural Networks, 1(4), 339-356. https://doi.org/10.1016/0893-6080(88)90007-x

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. ArXiv, abs/2303.17564.

Yu, K.K., Sturtevant, N.R., & Guzdial, M.J. (2020). What is a Quest? AIIDE Workshops.

Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2022). A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. ACM Computing Surveys, 56, 1 - 37.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. ArXiv, abs/1904.09675. [Original source: https://studycrumb.com/alphabetizer]