Master Thesis

Text generation for quests in multiplayer role-playing video games

Sander Koomen

Supervisors: dr. M. Theune
N. Bouali, MSc
dr. L. Gatti

November, 2023

Interaction Technology

Department of Human Media Interaction
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

This page intentionally left blank

Acknowledgements

First, I would like to thank my supervisors. A very special thanks to Mariët Theunen for her patience throughout this long process, for always making time for a healthy dose of constructive feedback over a cup of tea, and for the amount of freedom she has given me in this assignment. Another special thanks goes out to Lorenzo Gatti for his expertise and insights, and for his dedicated help when the technical implementation was at its most hopeless. Another special thanks to Nacir Bouali for his technical insights, for his patience during the registration process so that he could be involved in supervising this research, and for accepting to supervise a thesis on such a niche topic in the first place.

An additional thanks to Michael Bui for all of his insights on survey design and statistics, and his infinite patience for questions. Thank you to Femke Jansen for always believing in me and bringing out the best in me throughout our time together. Thank you to all of my friends for all of your input and support throughout this research process, particularly Daniela van Meggelen, Peter van der Burgt, and Maartje Huveneers. Your support means more to me than I can adequately put to words.

Finally, a last thank you to everyone else who has worked together with me in any way during my entire study period here. It is through these interactions that I have grown the most over these years.

Abstract

Large language models such as GPT-3 are increasingly popular for all kinds of language generation tasks. Additionally, video games are becoming increasingly expansive, and the required development efforts have increased over time. Interesting and immersive stories of high quality are hard to generate using traditional procedural generation techniques, and large language models could be used instead to generate stories in video games by picking up on the writing style of existing quests. This thesis was inspired by various previous research papers and theses on this topic, which generally fine-tuned older models such as GPT-2. This previous research highlighted a lack of suitable datasets and evaluation methods for video game quest story generation.

This thesis aims to continue this line of research by improving on limitations from previous research. This was done by gathering additional datasets and using these datasets to fine-tune newer models, with the aim of generating quest stories that fit better into the provided input context. It also proposes an evaluation framework for video game quest story generation, which is used to evaluate the generated results.

Four datasets were gathered from the multiplayer role-playing games World of Warcraft, The Lord of the Rings Online, Neverwinter, and The Elder Scrolls Online. These datasets contain thousands of quest entries, consisting of objectives and context information as input and quest descriptions as output.

These models were each used to fine-tune GPT-J-6B, an open-source large language model similar in performance to smaller versions of GPT-3. This resulted in four different models, namely one for each dataset. The descriptions generated with these fine-tuned models were evaluated by the researcher. Two of these models were then selected for evaluation with users. This was done through a survey and a focus group interview. Evaluation measures were set up for this survey. These are based on descriptive statements of the identified text quality factors of *Creativity*, *Consistency*, and *Fittingness*. Results from this survey were statistically analyzed using linear mixed models and showed only a small difference between the generated and original quest descriptions in an academic context.

The focus group interview allowed interested survey participants to expand upon their opinions on the storytelling mechanics in these games, and allowed them to discuss the original and generated descriptions in a side-by-side comparison. Results from the focus group interview showed that generated descriptions were more surprising and contained more mystery compared to original descriptions. However, many shortcomings were also identified during these evaluations.

The results from this research show an improvement over previous research in terms of generation quality, but the results still fall short in generating the correct background motivation for quests as this information is not present in the input. As such, additional research is required. This thesis provides an overview of the steps taken towards this purpose, and provides the datasets and code that was used for fine-tuning and generation. It concludes with recommendations for future work, which includes additional work toward model training, dataset structures, and evaluation methods.

Contents

Acknowledgements **i**

Abstract **ii**

Contents **iii**

Acronyms **v**

1 Introduction **1**
1.1 Open issues 2
1.2 Research questions 3

2 Background & related work **4**
2.1 Quests in video games 4
2.2 Quest generation 5
2.3 Language models 5
2.4 Creative text generation with language models 9
2.5 Meta-information 11
2.6 Datasets 12
2.7 Evaluation of computational creativity 13
2.8 Conclusion 15

3 Datasets **16**
3.1 World of Warcraft 16
3.2 Lord of the Rings Online 19
3.3 Neverwinter 21
3.4 The Elder Scrolls Online 22
3.5 Dataset comparisons 23

4 Fine-tuning and generation **27**
4.1 Methodology 27
4.2 Results 29
4.2.1 World of Warcraft 31
4.2.2 The Lord of the Rings Online 38
4.2.3 Neverwinter 44
4.2.4 The Elder Scrolls Online 48
4.3 Discussion 52

5 Quantitative evaluation **56**
5.1 Experiment design 56
5.2 Results 58
5.2.1 Participants 58
5.2.2 Methodology evaluation 59
5.2.3 Linear mixed-effects model 62
5.2.4 Participant comments 64
5.3 Discussion 64

6 Qualitative Evaluation **67**
6.1 Methodology 67

6.2 Results . . . 68
6.2.1 Open discussion . . . 68
6.2.2 Quest comparisons . . . 70
6.3 Discussion . . . 71

7 Conclusion & Future work **73**
7.1 Summary . . . 73
7.2 Research questions . . . 75
7.3 Limitations and shortcomings . . . 77
7.4 Future work . . . 77
7.5 Contributions . . . 79

References **80**

A Consent forms **84**

B Construct Q-Q plots and residuals plots **86**

Acronyms

LLM Large Language Model: a machine learning model trained on a large text dataset that can be used for a variety of language generation tasks.

RPG Role-playing game: a type of (video) game where players take on the role of a character in a fictional setting and experience the stories in this setting from the character's perspective.

1 Introduction

In the past few years, the ability of machine learning systems to generate coherent texts has rapidly improved. Increases in available computing power and the introduction of more efficient architectures such as the Transformer [1] have allowed the development of increasingly larger and complex language generation models. Because of these advances, machine learning models are trained on a large amount of examples of inputs with corresponding outputs to pick up on the information and structure in this data. More specifically, these large language models (LLMs) are trained on large text datasets in a self-supervised manner, with the purpose of informing the LLM on the structure and information present in this dataset. After training, a user can supply a piece of text as input, for which the model will estimate an appropriate continuation. It will append this output text to the original input text, and use this result as input again, and attempt to generate a continuation again. It will do this until it estimates the resulting output is complete. The performance of LLMs has been shown to improve by increasing the number of parameters and the amount of data used for training [2].

An example of the advancements in this field can be seen in the high development pace of LLMs such as OpenAI's Generative Pre-trained Transformer (or GPT) family of models. GPT-1 was released in 2018 and consisted of 117 million parameters [3]. It was trained on about 4.5 gigabytes of text data. Comparatively, the largest version of GPT-3 released in 2020 consisted of 175 billion parameters and was trained on over 500 gigabytes of text data [4].

LLMs can be used for almost any language task with a distinct input and associated output. This is also known as sequence-to-sequence (or seq2seq). Examples are translation, summarization, sentiment analysis, dialogue, rewriting, and code writing. If better performance on a specific task is desired, a model can be trained further on a tailor-made dataset containing examples of this task. This process is called fine-tuning. Because of their increased capability and applicability and the fact that language generation is very hard to set up manually, the use of LLMs is being researched in various contexts [5]. One of these contexts is computational creativity, which focuses on the application of text generation in creative fields. One of these applications is writing coherent and interesting stories, for instance for use in video games.

As the computational capabilities of computers continue to advance, so too do the possibilities for increasing the complexity of video games. New video games are becoming increasingly larger in scope, for instance by including larger game worlds, more expansive gameplay systems, and rich, interactive stories. This also means more extensive efforts from game developers. The quality of the story aspect of games is considered to be especially important in role-playing video games (or RPGs). These games allow players to take on the role of a fictional character and experience the game world and its stories through this chosen character.

In particular, multiplayer RPGs have evolved from their text-based origins into immense game worlds containing thousands of different stories in the form of quests. These quests are presented as assignments that fit into the game's world context or lore and provide the player with some sort of reward. Although the assignments used for quests in these multiplayer games can be quite straightforward, the variation present in their backstories is not. Large, interconnected worlds such as these require a consistent story as a setting for all of these quests, and game developers may have entire teams working on writing these quests.

As these games become increasingly sophisticated, so too does the expectation of players regarding the richness and depth of their in-game experiences. While procedurally generated content has been a staple in gaming [6], generating coherent and immersive stories that are constrained to a specific context has remained an open challenge. LLMs might be able to capture this rich and diverse context, and provide interesting, engaging, and possibly even adaptive quest backstories for players to enjoy. Writing high-quality pieces of quest text that fit into the larger narrative of a game world may become easier for developers by using these text generation methods. Additional background as well as previous work on this topic is discussed further in Chapter 2.

1.1 Open issues

Not much research has been done in the domain of creative text generation for video game quest stories by fine-tuning an LLM with a quest dataset. However, discussing the previous work in this domain as seen in Chapter 2 showed some open issues that still require further research.

The training databases used to train LLMs are purposely built to perform well on multiple different forms of text generation to create a well-rounded text generation model. Although larger LLMs that are trained on more diverse datasets are getting better at a variety of tasks, these models do not perform well on very specific tasks. In this case, a smaller fine-tuned model likely performs better [7]. Therefore, additional high-quality datasets containing examples of quest texts are required to fine-tune these smaller LLMs. Although previous research has shown promising results, it was observed that these datasets generally lacked the contextual information required for generating coherent backstories in complex game worlds. Additional information can be used to find patterns in word use between the context and the associated description, which could subsequently improve generation results. This means more extensive quest datasets are needed to fine-tune LLMs on specific generation tasks such as this. The datasets gathered and used in this research are discussed in more detail in Chapter 3.

Most research on this specific topic [8, 9] uses older, easier-to-train models such as GPT-2, which was released in 2019. This is understandable as computational requirements for newer models are continually becoming higher, and the required computational power may not be readily available for researchers or game developers. However, several techniques can be used to reduce the computational requirement needed and be able to fine-tune newer and larger models. This means finding the right balance of model performance and computational requirements is necessary to find effective solutions for specific text generation tasks. The models considered for this research are discussed in Chapter 4.

It is hard to estimate the quality of creative text generation because of the subjective nature of creativity. User evaluations are needed to make qualitative claims about the obtained results. However, there are no set standards for evaluating creative text generation. As such, it is hard to directly compare the results of these works. Complex evaluation criteria such as creativity and coherence are often evaluated through a single survey question, and results from these surveys are often interpreted as is. This means more substantiated evaluation methods using inferential statistics could improve the strength of conclusions drawn from the obtained results. The methods used to evaluate the results from this research are presented in Chapter 5 and Chapter 6.

1.2 Research questions

The open issues on this topic are multifaceted, and not all of these issues can be addressed completely in this research due to its limited scope. By improving on several limitations identified in previous research, this research aims to generate quest descriptions using multiple datasets with more context information, and perform a mixed-method evaluation of their quality compared to the original human-written quest descriptions with users.

The research questions for this research were therefore formulated as follows:

RQ1. To what extent can large language models be fine-tuned to generate backstories for multiplayer role-playing video game quests similar in quality to human-written quests?

RQ2. To what extent does the embedding of additional quest information influence the quality of quest story generation?

RQ3. What is the influence of the amount of quests used in fine-tuning on quest story generation quality?

This first research question was the main topic of interest, with research questions two and three providing additional insights when making direct comparisons. By trying to answer these questions, this research aimed to provide insights into the current possibilities of generating stories in the form of descriptions for quests in video games while highlighting new-found limitations and proposing avenues for future research.

2 Background & related work

This chapter introduces several topics related to text generation for video game quests. Firstly, video game quests and their current implementation in video games are explained, along with an overview of some of the different methods for generating text in a creative context. From there, the architecture and use of text generation models in creative text generation and more specifically, their use for generating video game stories are discussed. An overview of video game quest datasets that are currently available is also discussed. Finally, some of the methods used to evaluate computational creativity are discussed, and the entirety of the background is summarized.

2.1 Quests in video games

Many video games contain in-game assignments called quests. The structure of these quests in video games can vary wildly and as such, the definition of a quest is not formalized. By combining multiple previous studies, a generalized definition has been proposed by Yu et al. [10] as such: a quest is a set of tasks that a player must complete to get a reward, with the potential of having multiple different ways to complete the given task. In most games, quests are used by game designers to send players on their way into the game world to fight enemies, gather items, or help non-player characters and, by doing so, progress through the game. In turn, the game usually rewards these players with some form of in-game or meta progression. These rewards can vary wildly and could mean a continuation of the story, in-game currency, items or attributes that increase the player character's capabilities or that are purely cosmetic, or even just information on the game's lore. Quests are a part of both single-player and multiplayer games, and are found primarily in role-playing games (RPGs). Quests are generally the main mechanism for narrative progression in RPGs [11], and quest design is strongly related to other areas of game design such as level design and gameplay design.

Games use quests for different purposes and as such, quests are structured differently per game [12]. In general, single-player games generally present a single main storyline with perhaps some additional optional storylines and therefore tend to have relatively few quests in total. However, these quests are generally expected to be high quality [13] and somewhat complex, especially those related to the game's main storyline. Examples of this are The Witcher 3 or Fallout 3, where a single quest can have multiple intermediate steps and/or optional objectives and may allow for different approaches from the player. In comparison, multiplayer RPGs such as World of Warcraft or The Lord of the Rings Online often contain immense game worlds with many locations and factions with conflicts and subsequently, these games contain many different storylines emerging from these conflicts. Although some quests exist to motivate players to explore, most of these quests are constrained to a small portion of the entire game world. Due to the large quantity of smaller stories, these games generally use simpler quest structures [14], where the quests are not necessarily used to move a larger overarching story forward, but more so serve as a way for players to progress their characters with experience points, items, and game knowledge. Additionally, multiplayer games often use a series of small quests to keep the player interested and rewarded over time, instead of having only one big reward at the end of a long singular quest like in most single-player games. Multiplayer games usually also need a consistent narrative for all players to follow and for future updates to work upon, and can therefore not have too many quests with branching stories. This does not necessarily mean the writing for these quests is worse or that these structures apply to every game of this type, but that in general, the quests in these different types of games have varying purposes and structures.

2.2 Quest generation

Designing and writing these quests is a monumental task, both for the quality-oriented single-player games and the quantity-oriented multiplayer games. Since a quest's story is dependent on the quest objectives, and both of these are dependent on the game's available systems for interaction and the game designer's purpose for a given quest, a lot of design work is required to implement a single quest, and even more to design an entire game full of them [15]. Therefore, identifying the design patterns on any of these levels and finding ways to automate this creation process has been a research topic of interest for over a decade, collected in various survey papers. Hendrikx et al. [6] gathered various papers on procedural content generation, and dedicated a section to story generation in particular. They confirm the challenge of procedurally generated content, as well as the potential need for a complex artificial intelligence system that takes into account the story's cohesion and is constrained to this particular game's lore to successfully implement such a generation system. Garbacea & Mei [16] explore neural text generation and confirm the need for better constraints and evaluation metrics. They also note a lack of annotated datasets for specific attributes desired in consistent and complex text generation tasks.

The most straightforward method of procedurally generating quests consists of creating random combinations of handmade scenarios and objectives to form quests. This method is most often seen in video games as this is relatively simple to implement as well as being precisely controllable during game development. Using this method, game designers can easily create many different combinations of premade quest requirements, objectives, rewards, and stories with a consistent quality. Examples of games that implement such a system for optional quests are Starbound[1] and the Radiant system from The Elder Scrolls V:Skyrim[2]. However, as these quest generation methods are just slightly different combinations of premade assets with some variability in context, numbers, and awards, players will likely see the formulaic nature of these stories quite quickly as they exhaust most of the scenarios. Better quest text generation would therefore have to involve ways to generate entirely new text within the context of the game while still being controllable by the game's developers.

2.3 Language models

The current state-of-the-art method of text generation is in the form of large language models (LLMs) [17]. Whereas previous models used computationally intensive methods such as convolutions or recurrences, LLMs are based on the simpler and more scalable transformer architecture [1]. Previous models such as recurrent neural networks or long short-term memory networks generated an output in a linear fashion, and were limited in the amount of relevant contextual information they could keep track of during the generation of longer texts. By using a method called attention, the transformer model can take into account all of the previously generated text.

[1] Starbound Wiki - Quests
[2] Elder Scrolls V: Skyrim Wiki - Radiant

Transformer architecture

A transformer generally consists of an encoder that turns an input into an intermediate representation, and a decoder, which transforms this intermediate representation into an output. Although encoder-decoder models exist, decoder-only models such as the GPT family of models [18, 4] have become very popular in the past few years. Decoder-only models work auto-regressively. This means such a model uses an entire input text for generating a single output token, which is then appended to the original input. The input with the appended output token is then reused as input for generating the next token. This entire process is repeated until the model estimates the output is done. The lack of encoders in this architecture means that only previous text can be taken into account for generation as opposed to being able to fill in masked words with surrounding text, but this has been shown to not be an issue. For generation tasks that do not have any context after the output word, this functionality is not required and as such, decoder-only models are more efficient for these tasks. Transformers combine several techniques to generate text, which will all be discussed briefly to get an overview of the way transformers work. An overview of the mechanisms in the proposed Transformer architecture can be found in Figure 1.

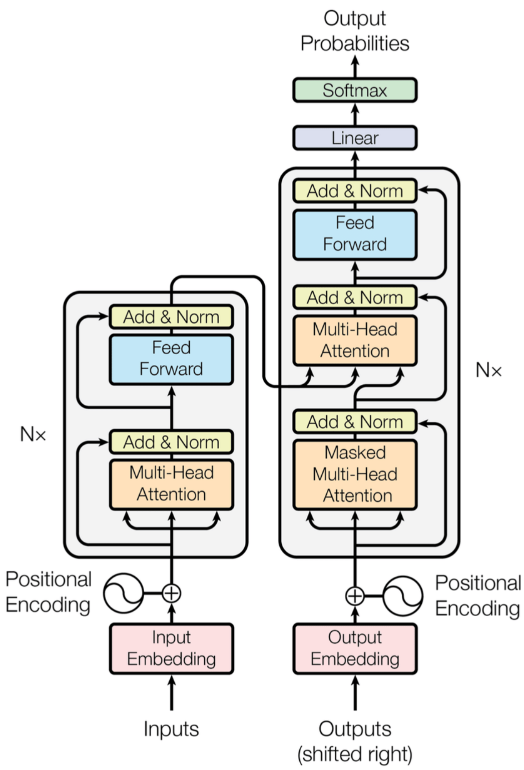

Figure 1: Transformer architecture with an encoder (left) and a decoder (right) [1]

Tokenization is used to turn all of the words or parts of words from the input into integers from the model vocabulary, since computers cannot work with text directly but need a numerical representation of words. Each of these integers has a related word embedding in the form of a vector. This vector contains all of the values for dimensions that the model identified during training. These are saved in an embedding matrix, where rows correspond to tokens, and columns correspond to dimensions. For an oversimplified example, observe the words 'Mouse', 'Fish', 'Elephant', and 'Whale' as plotted in Figure 2. It can be observed that the pairs 'Mouse'/'Fish' and 'Elephant'/'Whale' differ in their 'Location' value but coincide in their 'Size' value, while the pairs 'Mouse'/'Elephant' and 'Fish'/'Elephant' differ in their 'Size' value but coincide in their 'Location' value.

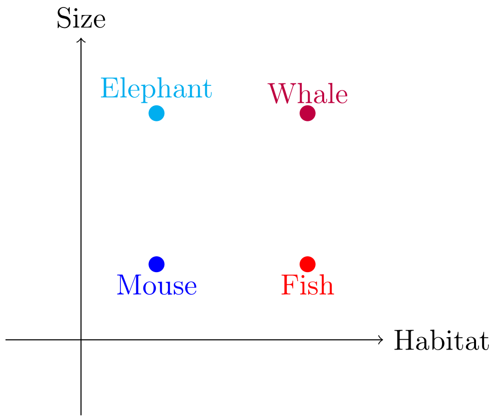

Figure 2: Word Embedding Example

This example contains two self-defined axes to serve for different dimensions, but for example the smallest version of GPT-2 has 768 dimensions, with a vocabulary of 50257 tokens. This means each of these tokens is measured and compared in 768 dimensions.

Since the size and content of the inputs used can vary considerably, the architecture should be able to take this into account. Attention is based on comparing all of the embeddings of tokens in the input with each other. Attention allows the model to learn to recognize the most relevant relationships between tokens in an input. This is done using vectors for a query (Q), key (K), and value (V). This could be compared to database retrieval, where a query vector is an entry that is compared to a key to retrieve an appropriate value. Attention is calculated using the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where Q, K, and V are the matrices formed using the word embedding vectors, and $\sqrt{d_k}$ is the dimension of the key (k) and query (q) vectors. In this formula, the attention weights are calculated in the dot product of the query and keys. This dot product measures the similarity between the query and key vectors for each pair of tokens in the input sequence. In essence, the higher this dot product, the higher the similarity between the query and the key. Each dot product between values in the query is compared to the values in the keys, which are then used to find a distribution of appropriate values. The division by the embedding dimension size prevents potential issues with very large values.

Self-attention keeps track of word relations within the input and output sentences by transforming their word embeddings with several linear operations. The last step is a softmax activation function. This produces a probability distribution of attention values for all of the input tokens that adds up to 1. Before the softmax function, the attention values can be positive, negative, or zero. The softmax function helps turn these scores into a set of values that represent how much attention should be given to each input token, with higher values indicating higher attention.

Multiple attention elements can be used simultaneously. This process is called multi-head attention, and it allows the model to keep track of multiple different interpretations of the same word but for different contexts. For instance, for the input sentence "I am counting on this board", several interpretations are possible. The word "board" in this sentence could be a piece of wood, or a whiteboard, or a board of directors, or the boarding of a plane. Additionally, "counting" could be interpreted as counting numbers or counting in the sense of relying on something. All of these could be valid interpretations, but if the model wants to correctly generate an additional token at the end of this sentence, it needs to know which interpretation is the most likely.

By calculating multiple attention scores using different weight matrices for calculating Q, K and V, multiple interpretations can be taken into account. The equation for multi-head attention is as follows:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2}$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

and the formula for attention is the same as seen in Equation 1. Each attention head independently calculates attention scores, and all of these outputs are then concatenated. Since these calculations are independent matrix multiplications, they can be executed in parallel which increases efficiency. A transformer model consists of many separate (multi-head) attention structures, all of which have different weights used for their calculations. The token embeddings as well as the values for the Q, K and V vectors are generally randomly initialized and will be updated during model training through backpropagation.

Training, fine-tuning and generation

The main goal of training is minimizing the difference between the model outputs and the desired outputs for a range of inputs. This is done by letting the difference between each of these outputs steer the model weights towards generating the desired output. This difference is generally called the loss. To avoid overfitting on a training dataset, a representative test dataset is used to calculate the loss on inputs and associated outputs that are not found in the training set. The goal is to have a low loss on both the training and test sets. Training is influenced by several parameters such as the learning rate (the magnitude of the differences made to the weights), batch size (the number of samples used for a single adjustment to the model weights), the amount of epochs (how many times the dataset is used to change the weights), and the optimizer (the methods used for calculating the loss and adapting the training rate) [19]. Additionally, techniques such as dropout [20] and weight decay [21] can be used as a method of making the model more robust and reduce the chance of overfitting.

After training on a large dataset, a transformer model has seen a lot of examples of text, and generally performs adequately on a variety of different generation tasks. These models can also be fine-tuned on a specific task by continuing training on a tailor-made dataset containing examples of this task. During this process of fine-tuning, the aim is to find a balance of information retention from the general dataset and task-specific performance. Another method for increasing generation performance that does not require additional training is in-context learning, which involves giving several example inputs and outputs alongside the actual input for which an output should be generated. This allows the model to pick up on the structure and content of the desired output.

The output of the language model is a probability score for each token in the vocabulary. In a greedy selection procedure, the token with the highest probability can be picked directly. However, more complex strategies can also be used to select a suitable token using sampling. Methods using sampling introduce variety in the output, which means one input can have multiple different outputs based on the sampling parameters used.

There are several parameters that influence the generation process, such as output length (the amount of tokens until an end-of-text token should be generated), temperature (a modulation on output confidence impacting text variety), word repetition penalty [22], n-grams (forbidding duplicates of successive words to reduce duplication) and many more, as well as parameters that influence the sampling strategy used for generation, such as top-k sampling (the number of highest probability vocabulary tokens to keep)[23], top-p sampling (the smallest set of token probabilities that add up to this number to keep)[24], typically sampling [25], or beam sampling (keeping track of multiple possible combinations of future tokens and picking the most likely).

Some downsides of this model architecture are the tendency of Transformer models to generate hallucinations or confabulations due to biases and errors in the training dataset and the stochastic nature of the sampling strategies [26]. Hallucinations can be intrinsic or extrinsic, where intrinsic hallucinations can be checked to be factually wrong when compared to the input text, and extrinsic hallucinations cannot directly be inferred from the input text.

2.4 Creative text generation with language models

Although a relatively niche topic in both natural language generation and story generation research [6], there is some relevant work related to video game quest generation using an LLM. This thesis is primarily based on the work of Van Stegeren & Myśliwiec [8], who collected World of Warcraft's quests and structured them in an XML-like manner. This was adapted from a method proposed by Zellers et al. [27], where opening and closing tags are proposed to label different input categories. In this method, only opening tags are used to indicate the location of different quest information. This was used to fine-tune the 774M parameter version of GPT-2 with the aim of generating quest descriptions from the combinations of quest titles, quest givers, and a quest objective. For example, an entry from their database may look something like this:

> <|startoftext|> Skirmish at Echo Ridge<|obj|>Kill 8 Kobold Laborers, then return to Marshal McBride at Northshire Abbey.<|text|>Your previous investigations are proof that the Echo Ridge Mine needs purging. Return to the mine and help clear it of kobolds. Waste no time. The longer the kobolds are left unmolested in the mine, the deeper a foothold they gain in Northshire.<|endoftext|>

The researchers evaluated their results in a survey by having participants rate 10 randomly selected original quests and 10 different randomly selected quests for which a description was generated using their fine-tuned model. This was done using five questions on a 7-point Likert scale. Three of these questions evaluated the description's creativity while the other two were related to language quality and coherence. While being evaluated slightly lower than their human-written counterparts, the quest generation system was deemed a success, albeit only for quests generated within the context and lore of World of Warcraft. Since the model was fine-tuned to this game specifically, it was not tested on generating quests not related to World of Warcraft.

A major downside to this method is that the researchers did not include information like the name of the quest giver or completer, and as such the quest generation sometimes referred to the quest giver (who would be "telling" the description to the player in the game) as if they were another person which would be incorrect. Although 20 quests is quite a lot of quests to compare, if these are not the same (or similar) quests it is also hard to directly compare their results. It is also not clear if the quests used to evaluate the model were left out of the training set, as this is not mentioned specifically in the paper. It is therefore not clear if the performance would be similar for unseen quests. The researchers conclude with possibilities for improvements, like the replacement of proper nouns with generalized tags to avoid factual or contextual errors, and using multiple video games to generalize the model's writing style.

The replacement of proper nouns with tags and sourcing from multiple video games was used in a thesis by Värtinen [9]. They discuss the simplicity of multiplayer quests, and focus on collecting quest data from several single-player video games instead. After replacing the proper nouns from all objectives and descriptions with placeholders, they use this database to fine-tune the 1.5B parameter version of GPT-2 to generate new quest descriptions. They explore a XML-like structure with tags similar to the one used by Van Stegeren & Myśliwiec [8] for fine-tuning and generation, but opt to go for a different data structure where the narrative of the quest is written out in simple sentences instead. This was done because of higher preliminary results. An example excerpt from their database looks like this:

> This is an RPG quest from a medieval fantasy video game. The questgiver is called questgiver. questgiver is a guard with a strong sense of justice. The questgiver gives a quest to the player. The player's objective is to kill character0. The player should first find kill character0 to complete their objective. This task can be completed in the following location: tasklocation0 (a sparkling fairy forest). The player will receive the following rewards for completing the quest objective: number0 platinum chips. The following characters are related to this quest: character0 (a male wizard and a dangerous lunatic). There are some important facts concerning this quest. character0 has killed several people before. This is the quest description, the questgiver explaining the quest to the player:

The resulting quality of the generated descriptions varied greatly, and although not many examples are given, the author notes questionable logic, poor grammar, unnecessary information, and repetition (for example: "I am Mogrul [...] My name is Mogrul. You might know me as Mogrul."). This could be because of the varied databases combined for fine-tuning, and the fact that this database is relatively small at 978 quests (compared to Van Stegeren & Myśliwiec's 24872 quests [8]). Previous work has shown that GPT-2-774M needs at least 1000 and preferably about 5000 text samples for adequate performance in creative tasks [28, 3]. These written statements also contain superfluous information like the first sentence in the example, which may lead the model to focus more on replicating this structure instead of learning the contents of the training set. Similarly, the used placeholders are not very distinct from previously seen tokens, and this may prove difficult for the model to differentiate. The research is concluded with a zero-shot experiment using such a prompt in a version of GPT-3 that is not fine-tuned. This experiment results in a generated quest description that is more coherent than the results from their fine-tuned version of GPT-2, showing the potential of in-context learning through zero-shot prompting and possibilities for fine-tuning tasks.

Ashby et al. [29] combine an LLM with a relational database containing information on entities from a game world and their relationships towards each other to generate new narrative structures for quests based on the current world state as well as input from the player. For instance, the player can (theoretically) type "I want to kill a dragon" in proximity of a non-player character, and this character will then try to create a quest from the relational data in the database. In the database, a dragon data entry might have a relation called "located_in" to a location data entry called "Poacher's Cave" which has a relation "located_in" to another location data entry called "Arelind", after which a sentence will be written such as "Fight the Dragon in the Poacher's Cave in Arelind". This is then fed into the language model to write a piece of contextual quest text such as "A mighty dragon resides in a cave in Arelind, to the west of here. Meet up with me there and we will fight the dragon together!" They fine-tune several differently sized variants of GPT-2 on the World of Warcraft dataset as set up by Van Stegeren & Myśliwiec [8] and use their own quest structure as input for these models. They evaluate their implementation through a survey where participants rate both their outputs as well as the original human-written quests and generated quests from Van Stegeren & Myśliwiec's model, as well through as a playtest that aims to evaluate player satisfaction between these outputs. From this, they find that the hand-written quests are better rated overall than any of the generated outputs, but that their generated quests have a relatively high dialogue relevance and a high participant satisfaction rate. Although some challenges are still unsolved such as obtaining high-quality user input, the researchers highlight the potential of improvements made to this system with further development in the form of player-driven quest generation. They note that the most common reasons for low-quality outputs are shortcomings of the training datasets (both their own relational dataset and the World of Warcraft dataset) such as spelling errors and size constraints, and hallucinations and biases from the model that may confuse the player.

2.5 Meta-information

Värtinen [9] also recognizes the need for more embedded quest information to aid the description generation, namely related locations, rewards, important facts, and related items, characters and factions. This is based on the work of Doran & Parberry [30], who analysed multiplayer video game quests and recognized a structure to quests based on the motivation of the quest giver. These motivations are Knowledge (gaining information), Comfort (improving living conditions), Reputation (doing something impressive), Serenity (keeping the peace), Protection (protecting from harm), Conquest (domination over others), Wealth (gaining valuable possessions), Ability (gaining or improve skills) and Equipment (gaining or trading items). Each of these motivations has several associated actions, for instance, the Knowledge motivation contains both spying and interviewing a character, as both actions give the player information. After the player has reported back, the quest giver's knowledge has increased. These actions can also be nested, as in order to retrieve an item, a player may have to interview a character before going to a location, killing an enemy, finding the item and returning it to the quest giver.

They continue to visualize these actions in a tree structure to show the possible actions taken by a player to complete a quest and subsequently implement this tree structure into a procedural quest generator. However, this was not yet evaluated. This has been expanded upon in several other works, namely Machado et al. [13] who added more actions to execute motivations to describe more complex quests from the single player RPG The Witcher 3. This work was then implemented into a quest generator for Conan Exiles, an open-world multiplayer game without quests of its own [31]. The addition of this to the game was evaluated to be equal in enjoyment and gameplay flow to the original game and deemed a success.

Another extension of Doran & Parberry's work is Breault et al. [32], who implement goals and motivations of quest givers into a state machine to generate quest narratives. They observe a quest motivation distribution in their generated quests similar to a collection of human-written quests. They also note that the complexity of the produced quests depends on the complexity of the world given as input.

Other common story structures can also be used for the generation of quests. De Lima et al. [33] use a three-act story structure to compare their quests to in order to achieve a stronger sense of tension in their genetic quest generator implemented into a zombie survival role-playing video game. Even though the amount of steps in their quests was limited due to time constraints in testing, quests were rated similarly to human-written ones. Ammanabrolu et al. [34] directly compare several Markov chain and neural network models for generation of cooking-related quests which consists of generating fitting ingredients for recipes that must be executed by the player in a text-based game. This is done by creating five different implementations: human-created quests, randomly generated ones, a simple and complex Markov chain model, and one neural network language model. These were all tested with humans on unpredictability, coherence, novelty and quality, and showed the language model performed better than both Markov chain models. In their results, they highlight the inverse relationship between coherence and creativity in generated stories, and the fact that longer and more complex stories are generally seen as more original. A similar contrast is found by Alvarez et al. [35], who analyze several video games for common literary tropes and use the observed structures to generate new possibilities for story structures. They also note that their measures for coherence and interestingness are inherently competing objectives.

2.6 Datasets

Fine-tuning a transformer model requires a sizable dataset of high-quality example inputs and outputs. The model can then be adapted to the dataset contents and structure with the aim of performing better on a specific language generation task.

Van Stegeren & Theune [36] note the lack of high-quality datasets and provide several options and considerations for future dataset gathering. These include gathering directly from the game files or accessing these files using developer-provided or community-provided modding tools, or scraping from user-generated resources such as fan-made wikis. Considerations that should be made for datasets are their diversity in contents, representativeness of the dataset with regards to high-quality writing, contextual richness of the information provided in the dataset, and accessibility of the data format used when making the dataset available for further use. They also provide several datasets with different types of flavor text from single-player RPGs. These are quest texts from Torchlight 2, dialogue texts from Star Wars: Knights of the old Republic, and texts from books found throughout the Elder Scrolls series. They also highlight the need for further research in the area of machine learning-supported text generation methods.

Van Stegeren & Myśliwiec [8] use a collection of quests from World of Warcraft. This dataset has also been made publicly available for further research and is sourced from the fan-made wiki WoWHead.[3] While this is a large dataset, it contains quests from the first version of the game released in 2004 to the Battle for Azeroth expansion released in 2018. In this period, the game has undergone many structural changes during which the original game world has been completely restructured and more complex quest mechanics have been implemented. Additionally, this game started including in-game cutscenes from the Wrath of the Lich King expansion onward and started expanding on this trend in the following expansions with out-of-game cutscenes, both of which mean not all of the story content, and with that the context for the game world, is provided by just the quests anymore. Similarly, placeholders for the player characters name, race or class that are normally depicted with <name>, <race> and <class> have been removed altogether, causing some poor sentence structures where these tags were previously used.

The database used by Värtinen [9] consists of several single-player RPGs, some of which are specifically known for their strong stories. These video games are Baldur's Gate 1 & 2, The Elder Scrolls 1 through 4, Torchlight 2 and Minecraft, for a total of 978 quests. These were scraped from fan websites, gathered using modding tools, or were previously collected by other researchers [36].

Although more game-specific quest datasets exist in the form of fan-made wiki pages, these are generally not easily queried. The text on these webpages often contain markup language for formatting of the text, making it harder to access the desired plaintext present on these pages. However, these might be collected through methods such as scraping, after which data on these pages can be cleaned and formatted for use in datasets.

2.7 Evaluation of computational creativity

There are several different measures commonly used in the evaluation of natural language generation (NLG) systems. This is partly because of the large number of tasks made possible with these types of systems. This includes translation of sentences, summarization of texts, dialogue with a chatbot, text rewriting, text style transfer, and many different forms of creative writing. Examples of creating writing include writing poems, jokes, puns, puzzles, song lyrics, and in this case, writing background stories from a quest description.

The evaluation of NLG systems is difficult because of these varied specific domains, and the fact that there is often not a single output, but a range of valid possibilities for the output. Metrics that measure the similarity between output and reference texts such as BLEU [37] and ROUGE [38] are often reported. Both BLEU and ROUGE do this by comparing the overlap of successive words (or n-grams) between the generated output and the validation text output. Similar metrics exist that use other text similarity calculations, such as edit distance and synonym matching, but these are used less frequently. While these measures are suitable for assessing the quality of natural language generation where the expected outputs do not contain a lot of variation or there is a set output value, it may not necessarily be good for evaluating creative systems as the outputs of these systems are much more varied. For instance, the translation of a sentence will likely have some variety in wording but most outputs will likely be quite similar. On the other hand, a joke generator is more open-ended and may generate wildly different jokes for a single input, which means there is no strict "best" output and it becomes much harder to estimate the output quality using these measures of similarity.

[3] WoWHead - World of Warcraft Database

Perplexity is another automatic measure that is used for assessing a language model's average uncertainty for the next token in the generation process based on the current output. The lower this measure is, the more certain a model is about each next token. This does not mean it is necessarily accurate, but that the model is certain about the specific prediction it makes instead of doubting between several options. The downsides of using this measure are that it requires a sizable unseen validation dataset separate from the training and test datasets used in training or fine-tuning the model, and since it requires test datasets, perplexity scores are not directly comparable between models. Instead, different models should measure perplexity on the same test datasets. These datasets should also be selected based on the task that the model is trained for.

Another common method of evaluation is with human participants. This can consist of expert annotations or evaluation surveys on measures such as creativity and text consistency. Van der Lee et al. [39] focus on evaluation with human participants and provide in-depth advice and best practices for the field of natural language generation. They note the shortcomings of automatic metrics such as BLEU, ROUGE, and METEOR, as these are generally under-informative and do not correlate well with human evaluations. Additionally, they note the lack of agreement on evaluation of language generation systems. They highlight several considerations for evaluation methods such as the need for clear evaluation goals, hypotheses, and measures, as well as outlining a range of best practices and potential pitfalls for both qualitative and quantitative evaluation methods. Some example recommendations are the use of inferential statistics such as mixed-effects models, and question order randomization in surveys.

Hämäläinen & Alnajjar [40] also focus on the evaluation of creative systems with human participants. They highlight the need for more consistent and higher-quality testing of creative systems. By comparing several papers, they find several common measures used to evaluate text quality, although the definitions and parameters for each measure differ a lot per paper. These measures include meaning, syntactic correctness, novelty, relevance, and emotional value. Evaluation is often done with surveys, generally with 5-point Likert scales, or by ranking the generated results by the participant's personal preference. They emphasize the need for targeted evaluation of the property that is modeled, more detailed responses from human participants, and proper reporting of demographics.

Jordanous [41] evaluates some of the common practices of evaluating creative systems and discusses the challenges with these methods. They emphasize the lack of standard, generalized computational creativity evaluation practices and provide a framework of questions and recommendations for best practices for this type of evaluation, along with identifying 14 components of creativity that can be used in this method. These components are Active Involvement and Persistence, Dealing with Uncertainty, Domain Competence, General Intellect, Generation of Results, Intention and Emotional Involvement, Originality, Progression and Development, Social Interaction and Communication, Spontaneity/Subconscious Processing, Thinking and Evaluation, Variety/Divergence/Experimentation, and Value. As not all computational creativity domains focus on the same topic or context, they note the importance of selecting the right creative properties that should be taken into account when designing these systems and when evaluating them afterward. Their standardized procedure for evaluating creative systems (SPECS) consists of a preliminary question list that aims to narrow down the most and least relevant creative properties of the system in question from the list of 14 components of creativity. Researchers can then use these relevant components to devise relevant tests for these properties to reduce duplicate work in defining creativity.

2.8 Conclusion

Quests are a storytelling tool unique to video games, and can provide a wide range of gameplay and story experiences. Quests can be described as little pieces of conflict that provide the players with challenges to overcome in order to get a reward. This is often used in RPGs as a way to keep the story moving forward. However, quests in multiplayer RPGs are often derided as busywork to keep the player engaged as they progress through a game's world to get to the most interesting and challenging parts at the end. However, quests in this type of game provide a unique opportunity for storytelling in a vast, interconnected, shared world populated by different players with differing views and goals. Designing and writing these quests for a game of this scale is an immense task for game developers, and making these quests more interesting and worthwhile to experience takes even more work. However, this might become easier through the use of writing tools such as large language models. Large language models are machine learning models that can generate new pieces of text based on an input text. Out of the box, these models are not necessarily usable for a highly specific task such as writing these quests, but by fine-tuning these models on tailor-made datasets focused on a specific task, their performance on this task can be increased greatly. However, both fine-tuning a language model and generating results from such a model require a relatively large amount of computational power and storage, as well as relatively large datasets. While language models are a promising avenue for creative text generation, most previous work on this topic was done using relatively small models such as GPT-2, and either used a dataset from a single game or a combination of datasets from multiple different games. There are not many available datasets, and the available ones generally lack the context that may be required for storytelling in large and interconnected worlds as present in these games. Datasets compiled from multiple games provide a broader range of training data, but this is done at the expense of outputs tailored to a game's writing style and context. This could be solved by replacing proper nouns with placeholders and filling these placeholders back in manually after generation. This does require more work from the user in preparing the training data beforehand and filling in the placeholders after generation. Although not many quest databases from multiplayer RPGs are currently readily available, these could be collected from various sources such as a game's original files or externally compiled sources such as player-curated wikis or tools. There are also challenges in defining a universal standard for evaluating the quality of the obtained results, and as such, most research attempts to define this for their specific application. There is no standard evaluation method for generated results, likely because of the wide range of possible applications for creative text writing. As such, automated metrics often do not translate well to these domains either as these require a baseline dataset to compare generation results to. Quantitative evaluation is generally not well defined, and results are often interpreted as is instead of using statistical methods to obtain more substantiated results.

3 Datasets

This chapter shows the actions performed to select, gather, and clean the datasets collected for fine-tuning. Lastly, the differences between the datasets as well as any other peculiarities or challenges with these datasets are discussed.

Requirements were set up based on the insights gained from previous research. Selected datasets should have a clear objective (i.e. the assignment) and an associated description (i.e. the backstory) for each quest. Additional context about the quest (or meta-information) such as the title, relevant locations and factions, and relevant non-player character names should also be collected alongside the objectives and descriptions, as this contextual information may further inform the LLM on quest structures and word use. This may prevent the LLM from hallucinating incorrect contextual information when generating the quest description. Based on the fact that the preferred size of a dataset used for fine-tuning GPT-2 was already estimated to be 1000 entries and the fact that newer and larger models will be used here, the gathered datasets should contain at minimum 1000 and preferably at least 5000 entries. Although multiplayer RPGs are more likely to fulfill this requirement as these generally contain more quests, single-player games are not specifically excluded from this search. Finally, the selected games were also required to have a depository of quests publicly available, such as a dedicated wiki that contained the required quest information.

Four video games were found that fulfill all of these requirements: World of Warcraft, The Lord of the Rings Online, Neverwinter, and The Elder Scrolls Online. The steps taken to gather a dataset for each of these games are discussed below.

3.1 World of Warcraft

World of Warcraft (WoW) is a multiplayer online RPG originally released in 2004. It takes place in the Warcraft universe, which consisted mostly of strategy games before the release of World of Warcraft. The game has two major opposing factions that players can choose to join. These are the Alliance and the Horde, and players in one faction can attack players of the opposing faction in the game world. As such, this dataset contains a lot of similar versions of quests for both of these factions so that players from these two factions would encounter each other in the game world and come into conflict.

Collection

The database was scraped from a World of Warcraft wiki[4] that consists of data originally gathered from the game files using a database tool called AoWoW.[5] The dataset was created by looping through the quest ID in the URL and scraping the same data elements from each page using regular expressions in a scraping tool called WebHarvy[6]. From there, a CSV file was exported. Further sanitizing of the dataset was then performed in Microsoft Excel. This World of Warcraft dataset was specifically chosen to only include the first two expansions. This was done because the third expansion (titled Cataclysm) significantly changes the game world layout and therefore also changed or removed many of the original quests. As such, the quest context in a dataset with all of these quests would make the contextual meta-information such as locations significantly more complex. As it was not

[4] EvoWoW WotLK - World of Warcraft Wiki
[5] Github - AoWoW Database Viewer
[6] WebHarvy - Web scraping software

possible to include the expansion for each quest during gathering, the choice was made to only focus on quests that existed at the same time in the game world to maintain story consistency in the dataset.

The final dataset contains 6892 entries. An entry consists of 594 characters on average (without tags), of which the description is on average 404 characters long. A typical example of an entry from this dataset can be found in Table 1, which contains the properties, associated tags, and an example entry.

Table 1: World of Warcraft - Example Quest

<table>
<tr><th>Property</th><th>Tag</th><th>Example value</th></tr>
<tr><td>Title</td><td><|startoftext|></td><td>Syndicate Assassins</td></tr>
<tr><td>Level</td><td><|level|></td><td>33</td></tr>
<tr><td>Race</td><td><|race|></td><td>Alliance</td></tr>
<tr><td>Category</td><td><|category|></td><td>Alterac Mountains</td></tr>
<tr><td>Quest giver</td><td><|quest_giver|></td><td>Magistrate Henry Maleb</td></tr>
<tr><td>Quest completer</td><td><|quest_ender|></td><td>Magistrate Henry Maleb</td></tr>
<tr><td>Quest line</td><td><|quest_line|></td><td></td></tr>
<tr><td>Required quests</td><td><|required_quests|></td><td></td></tr>
<tr><td>Unlocked quests</td><td><|unlocked_quests|></td><td></td></tr>
<tr><td>Objective</td><td><|objective|></td><td>Kill 12 Syndicate Footpads and 8 Syndicate Thieves, then return to Magistrate Henry Maleb in Southshore.</td></tr>
<tr><td>Description</td><td><|description|></td><td>I am the new magistrate of Southshore, recently assigned after the assassination of the previous magistrate. The assassins were never found, but through our investigations we're almost certain they were hired by the Syndicate - a group of thieves led by villainous nobles of the now fallen kingdom of Alterac. The Syndicate has a camp in Sofera's Naze and Corahn's Dagger, north and west of the Horde-occupied Tarren Mill. Slay the Syndicate members you find in these camps.</td></tr>
</table>

Cleaning

Entries that were either completely empty or were missing important parts such as a title, quest giver/completer, objective or description were removed. Disabled or unused quests as well as quests with placeholder titles such as "<TEST> HEY MISTER WILSON!" were removed. To avoid redundancy and a sparse dataset, several columns of meta-information were combined into single columns. The first case of this regards a quest's character level requirement. Some quests contain a minimum character level requirement while other quests contain a "recommended" character level at which the quest is recommended to be played. As these were not always present for every quest, these columns were combined to use the recommended level unless it was absent, in which case the minimum level requirement was used. The second case of combining meta-information was for the *Category* column. Previously, these quests contained columns with the character class for which the quest was available, the relevant sub-faction for which the quest took place, and the relevant profession to which the quest was constricted. However, only one of these columns was ever filled in for each quest, as there were no quests that combined multiple of these factors as a requirement. Therefore, these columns were combined into the category column, which now contained a single relevant property to the quest based on the available information. If none of these were available, the region in which the quest takes place was used instead, as this meant the quest was likely independent of these factors and simply available for every player in this region of the game world, making the region the most relevant property. Additionally, because groups of quests that are linked by an overarching story (or questlines) don't have a title, the first quest in a questline is also used as the title for the entire questline. Finally, quests that did not have a clear objective or description were removed. The same was done for some quests that contained a lot of periods or dashes. Take for example the following excluded description:

> I have not long to live... so... so... weak. ...the b-b-brave dwarf, Lonebrow, has been sent to w-w-warn Falfindel... But before I... d-d-die... ...my wife, Treshala... her pendant of bonding to me... stolen... by one of the f-f-foul aggressors... Find it p-p-please... and return it to Treshala in Darnassus... ...along with w-w-word of my lonely... death.

Although this is an interesting description, there are a lot of periods and dashes present in this text and these are used quite randomly. Using this quest as a training example could mean an increase in the chance of a period or dash becoming a valid choice for the next token in generation at any point. As random periods or dashes in an otherwise normal text would likely reduce the quality of the output and editing the quest to remove these punctuation marks would still lead to texts containing incorrect grammar, quests like these were excluded.

3.2 Lord of the Rings Online

The Lord of the Rings Online (LotRO) is a multiplayer online RPG originally released in 2007. It is set in J.R.R. Tolkien's Middle Earth and allows the player to play through a story set alongside the original story from the Lord of the Rings books.

Collection

The Lord of the Rings Online dataset was gathered from the LOTRO companion application[7], a fan-made tool that sources its data from the game files. The final dataset contains 7411 entries. An entry consists of 689 characters (without tags) on average, of which the description is on average 477 characters long. A typical example of an entry from this dataset can be found in Table 2, which contains the properties, associated tags, and an example entry.

Table 2: The Lord of the Rings Online - Example Quest

<table>
<tr><th>Property</th><th>Tag</th><th>Example value</th></tr>
<tr><td>Title</td><td><|startoftext|></td><td>A Call to Elves</td></tr>
<tr><td>Level</td><td><|level|></td><td>30</td></tr>
<tr><td>Category</td><td><|category|></td><td>Volume I, Book 3: The Council of the North</td></tr>
<tr><td>Quest giver</td><td><|quest_giver|></td><td>Halbarad</td></tr>
<tr><td>Quest completer</td><td><|quest_ender|></td><td>Gildor Inglorion</td></tr>
<tr><td>Previous quest</td><td><|prev_quest|></td><td>Fallen Once More</td></tr>
<tr><td>Next quest</td><td><|next_quest|></td><td>Tending the Glade</td></tr>
<tr><td>Objective</td><td><|objective|></td><td>To deal with the threat from Angmar, Halbarad has decided that he must call a council of the Free Peoples of the North Downs – the Council of Esteldín.</td></tr>
<tr><td>Description</td><td><|description|></td><td>To the south of Esteldín, there stands a small refuge of the Elves. Many long years have the Elves dwelt there; however, in recent days, most of the Elves left the glade. What is worse, I have heard that those few that remained were slain by Stone-trolls from the North. It is an irony that Gildor Inglorion, an Elf-lord of Rivendell, was coming to give word of the Enemy's movements, but arrived too late to save those who remained. He should be called to the Council, but I fear that he will be too consumed by the desire to protect both the glade and his people from their rage. It will be your task to go to Lin Giliath and convince Gildor to come to the Council.</td></tr>
</table>

[7] Github - LOTRO Companion Quest Database

Cleaning

Although most quests have multiple lines of dialogue and can contain an objective that changes over time, the description is generally only related to the first of these objectives. As such, only the first objective for each quest is used in the dataset. A lot of quest titles contained a certain Volume/Book/Epic/Intro/Prologue that the quest was from, which has been moved from the title to the category. Several quests contained very similar objectives and descriptions, so these were combined into a single quest. This was mostly the case for repeatable quests, where there was only a small difference in wording. Temporarily available event quests were also removed, as the story in these event quests was generally much shorter and not as interesting as the normal quests.

A possible limitation of this dataset is the fact that some quests originally contained multiple objectives that need to be completed in order, but only the first objective was used in this dataset. This may lead to some confusion with including the name of the quest completer in the description, as originally it might have taken several steps to get to them. Since the model cannot know this, the generated description may already mention the quest completer in the description, while the original description will not.

Since quotation marks were used inconsistently at the start of sentences and were removed along with newline characters, it is no longer exactly clear which parts of the text are spoken words and which parts are actions taken by the character. Take for example the following excerpt:

> I am Golodhril. My sister Sadorwen and I have come to see the fabled Forest Beneath the Mountain! She sighs and slumps. Sadly, we have seen barely anything save rocks and weeds.

Since this differentiation between character speech and character actions is not clearly shown in the text, this should be taken into account as a limitation of the dataset. The placeholders for the player name, class, and race were inconsistent with the other datasets and were therefore replaced with <name>, <class>, and <race> respectively.

3.3 Neverwinter

Neverwinter is a multiplayer online RPG originally released in 2013. It is based on the popular campaign setting "Forgotten Realms" from the tabletop game Dungeons & Dragons. The game currently has 27 expansion modules that provide additional storylines for players to experience.

Collection

The Neverwinter dataset was scraped from an XML export of all pages that fall under the wiki category "Quests" on the Neverwinter Fandom wiki.[8] The desired properties were then extracted from this XML file using regular expressions in Python and saved as a CSV plaintext file. Further sanitizing of the dataset was performed in Microsoft Excel.

The final dataset contains 1257 entries. An average entry consists of 568 characters (without tags), of which the description is on average 399 characters long. A typical example of an entry from this dataset can be found in Table 3, which contains the properties, associated tags and an example from an entry.

Table 3: Neverwinter - Example Quest

<table>
<tr><th>Property</th><th>Tag</th><th>Example value</th></tr>
<tr><td>Title</td><td><|startoftext|></td><td>Hunting the Hunters</td></tr>
<tr><td>Level</td><td><|level|></td><td>23</td></tr>
<tr><td>Quest giver</td><td><|quest_giver|></td><td>Sergeant Yates</td></tr>
<tr><td>Quest giver location</td><td><|quest_giver_loc|></td><td>Blackdagger Ruins</td></tr>
<tr><td>Quest giver</td><td><|quest_ender|></td><td>Sergeant Yates</td></tr>
<tr><td>Quest giver location</td><td><|quest_ender_loc|></td><td>Blackdagger Ruins</td></tr>
<tr><td>Previous quest</td><td><|prev_quest|></td><td></td></tr>
<tr><td>Next quest</td><td><|next_quest|></td><td></td></tr>
<tr><td>Objective</td><td><|objective|></td><td>Hunt down and slay Bandit Huntsman in the mines area, and across all of Blackdagger Ruins.</td></tr>
<tr><td>Description</td><td><|description|></td><td>I could really use the help of a <class> of your stature. We are holding our own here in the camp, but any time we venture out bandit huntsman start tracking and shooting us. It would be a great help if you could go and clear them out. Let them know who's in control of the highroad!</td></tr>
</table>

[8] Neverwinter Fandom wiki

Cleaning

Quests that had no description or objective were removed. Some optional quests are repeatable and contain different texts between the first time the quest is received and any of the following times it is received. Only the text for the first time the quest is received was used for this dataset as this generally gives a better explanation for the context of the quest. Quests that mention meta-elements such as premium currency, out-of-universe characters like "Rewards Claim Agent", or skill points were removed. Quests that mention the fact that they can be completed daily or mention daily rewards were also removed. Markup language present in the text that was used by the wiki to link to items or locations have been removed. The player name placeholder <character name> was changed to <name> to be consistent with the other datasets.

3.4 The Elder Scrolls Online

The Elder Scrolls Online (ESO) is a multiplayer online RPG originally released in 2014. It is part of The Elder Scrolls game series which mostly consists of single-player RPGs. Where most of these older games were set in a single part of the continent called Tamriel, The Elder Scrolls Online combines many of these game zones into a single game world. Quests in this game are generally quite long and can have multiple steps. Quests are generally presented through spoken dialogue, but a starting objective text and a description used in the quest log are still given as text resources to the player, and as such, this game was still selected. This does however mean that the objectives are relatively simple and do not contain a lot of context that is explained in the description. This lack of a direct relation between objective and description may be more challenging for the LLM to find patterns in compared to the other datasets.

Collection

The Elder Scrolls Online dataset was scraped from an XML export of all pages that fall under the category "Quests" from The Elder Scrolls Online on the Unofficial Elder Scrolls wiki.[9] This wiki is created and maintained by players, using player-made tools to gather quest information from the game files. The desired quest properties were extracted from the exported XML file using regular expressions in Python and saved as a CSV plaintext file. Further sanitizing of the dataset was performed in Microsoft Excel, after which another CSV was created with the tags inserted as seen below.

The final dataset contains 2021 entries. An entry consists of 280 characters (without tags) on average, of which the description is on average 159 characters long. A typical example of an entry from this dataset can be found in Table 4, which contains the properties, associated tags, and an example from an entry.

Cleaning

Some quest objective-description pairs differ depending on which quest giver the player receives the quest from, or whether or not they have completed the quest before. These have differing objectives and descriptions, but both of these are present in the same quest entry. The choice was made to reduce these double entries to a single objective and description per quest with a preference for the first time a quest was received, instead of having two entries for the same quest title but with different objectives and descriptions.

[9] Unofficial Elder Scrolls Wiki

Table 4: The Elder Scrolls Online - Example Quest

<table>
<tr><th>Property</th><th>Tag</th><th>Example value</th></tr>
<tr><td>Title</td><td><|startoftext|></td><td>Badwater Mine</td></tr>
<tr><td>Level</td><td><|level|></td><td>31</td></tr>
<tr><td>Quest type</td><td><|quest_type|></td><td>Side</td></tr>
<tr><td>Faction</td><td><|faction|></td><td></td></tr>
<tr><td>Location</td><td><|location|></td><td>Alik'r Desert</td></tr>
<tr><td>Previous quest</td><td><|previous_quest|></td><td></td></tr>
<tr><td>Next quest</td><td><|next_quest|></td><td></td></tr>
<tr><td>Objective</td><td><|objective|></td><td>Rescue a miner's partner from a collapsing mine.</td></tr>
<tr><td>Description</td><td><|description|></td><td>I met a miner named Samsi at the Badwater mine. She tells me that the mine is collapsing and her partner is trapped inside. I told her I would go in and find her partner. If he's dead, she still needs his supply pack.</td></tr>
</table>

Quests that contained multiple quest givers and/or multiple locations to receive the quest were reduced to a single quest giver and quest location respectively. For some quests, multiple quest givers were listed but only one description was given. In these cases, the correct quest giver for these objectives was selected by using the information present on the wiki. Repeatable quests that are specifically mentioned to be daily/weekly/monthly quests have been removed, as these quests were not deemed to be interesting from a storytelling perspective. The same was true for temporarily available event quests as the descriptions for these are substantially shorter and were deemed to be less interesting than the normal quests. Duplicate quests have also been reduced to a single quest.

3.5 Dataset comparisons

While each quest from these datasets contains at least a title, quest giver name, objective, and description, several differences could be observed between these datasets. These datasets were made public for potential further use in research. However, the information in the quest databases cannot be assumed to be completely perfect, and some errors or mistakes may have been missed by the researcher because of the subjective nature and time constraints of cleaning the datasets. Additionally, these datasets contain copyrighted material and should only be used for research purposes.

Size

The most obvious difference between these datasets is their size. While the World of Warcraft and The Lord of the Rings Online datasets contain around 6800 to 7400 entries respectively, the Neverwinter and The Elder Scrolls Online datasets only contain around 1200 to 2000. Similarly, the average length of entries for The Elder Scrolls Online is significantly shorter than the other datasets, being about half the total length per entry. Although many more expansions are available for the World of Warcraft dataset, this subset of the first two expansions was specifically chosen to maintain game world consistency within the dataset, which means that the database consists of "only" about 7000 quests. Comparatively, adding up all quests from all expansions nets around 35000 quests. Adding an expansion column to this more expansive quest database could then prevent confusion about the context of the quest and maintain text generation quality.

Writing style

The use of the current World of Warcraft dataset may be more challenging for the LLM to fine-tune on because it contains multiple writing styles. For instance, the language use between different races in the game world differs greatly, where for instance Trolls speak with varying Caribbean and African accents, whereas Dwarves speak with Scottish accents, and Goblins talk with a New York or New Jersey accent. Since these are still interspersed with normal English words, the LLM could have trouble differing this word use from the General American writing style and word use that most other characters use. The column for quest giver names could already be somewhat helpful for this as there are some standard naming conventions that could be picked up by the LLM. However, adding another column for quest giver race could improve on this aspect even more. This was unfortunately not easily done at the time of gathering the datasets because the wiki pages do not list the quest giver race directly on the quest's page or even the quest giver's dedicated page. Instead, this is only shown in the quest giver's related audio files. The variety of language between races is most pronounced in the World of Warcraft dataset but is true to a certain extent for the other datasets as well. Similarly, among the different races and factions, there are many more sub-factions with varying levels of hostility towards different player races, further complicating the way certain races and factions are mentioned in the quest descriptions. This level of variety may also make the dataset more valuable, depending on how well the language model can adapt to this property. The World of Warcraft dataset also contains many quests with the same title for quests in a questline. This could be compared to the quests of The Elder Scrolls Online or The Lord of the Rings with multiple objectives, except that these quests are separated and have distinct objectives and descriptions.

While quests in The Elder Scrolls Online dataset contain an objective and a description, the descriptions are all written in the first person as entries in the player character's journal. Comparatively, the other datasets instead use the description as a piece of expository monologue from the quest giver. Many of the quests in The Elder Scrolls Online also contain multiple steps, which would mostly be split up into smaller quests that form a quest line in the other datasets. Since the quests also contain a lot of voiced dialogue instead of text, not all of the context for the quest may be clear from just this information. Similarly, objectives in the Neverwinter dataset are quite short and worded in an imperative way: instead of "Character X wants you to investigate Y to do Z", it is worded as "Investigate Y", which decreases the amount of information available for the LLM.

Character actions

There are some peculiarities and exceptions in these datasets that may influence the generation quality. For instance, most of these datasets contain a distinction between speech and actions performed by the non-player character. These actions are not visually performed by the non-player character but are instead provided in textual form to add to the immersion of the description. As shown in Chapter 3.2, the difference in quest giver speech and actions is not properly distinguished in The Lord of the Rings Online dataset. This is because quotation marks were used only at the start of sentences in the descriptions. Since whitespace between paragraphs of quest descriptions could not be taken into account in the plaintext data format used, it was no longer clear which sentences were originally used as character actions, and which sentences were part of speech. As such, these quotation marks were removed.

Both the Neverwinter and World of Warcraft datasets suffer from this somewhat as well, as some quests are not given by non-player characters but are instead provided by inventory items or objects in the game world (such as machines) that the player can find in the game world. In these cases, there is no quest giver character that speaks to the player character directly. Instead, a second-person storyteller perspective is chosen to convey the quest text for this type of quest, and different ways of denoting actions and dialogue is used. Take for example the following description from the World of Warcraft dataset:

> The object is not a crystal at all, but some form of organ, encased in ice. As you touch it, your hand reels in horrible pain. You feel as if your hand has been cut, yet there is no visible wound. As you collect your thoughts, you hear a young boy's voice nearby. "You really shouldn't have done that."

In this example, the dialogue is surrounded by quotes, and the player actions are now denoted as the "normal" text. This is in contrast to quests given by non-player characters in World of Warcraft. In those quests, actions performed by the non-player character are instead placed within angled brackets and their dialogue is the "normal" text. This can be seen in the following description:

> <Barthus tears out the last page and rolls it up.> Take this to Hellscream at once! In the meantime, I'm going to send a team to scout out the farms. Be sure to tell Hellscream that we're going to need reinforcements.

Additionally, many objectives in the Lord of the Rings Online dataset, especially those that relate to story quests, are not really objectives at all but are more so summaries of the story events taking place. Take for example the following objective from The Lord of the Rings Online dataset:

> Gandalf has come to Skarháld with a small party of dwarves including Hrostyr, the personal guard of King Thorin.

Properties

See Table 5 for a comparative overview of the contextual information present in each dataset. Attributes that contain an asterisk (*) are covered in the *Category* attribute of that dataset. This is because of the differences in the way each dataset used this *Category* attribute.

As mentioned in the World of Warcraft section, some attributes that could further improve the results would be tags for different expansions and/or periods of the game's development. This could inform the LLM about possible differences in writing style and game world context. Although map markers are also used in these games to convey location information, these are not used as the main information source for directing the player and are unlikely to be helpful when added as meta-information.

Table 5: Attribute overview per game

Game / **Attribute**	World of Warcraft	The Lord of the Rings Online	Neverwinter	The Elder Scrolls Online
Title	X	X	X	X
Level (range)	X	X	X	X
Race	X			
Faction	*			X
Category	X	X		
Type				X
Giver name	X	X	X	X
Giver location			X	X
Completer name	X	X	X	
Completer location			X	
Quest line	X	*		
Prerequisite quest	X			
Previous quest		X	X	X
Unlocked quest	X			
Next quest		X	X	X
Objective	X	X	X	X
Description	X	X	X	X

[*] Included in *Category* attribute

4 Fine-tuning and generation

This chapter will discuss the process of selecting an appropriate language model for this research, as well as the process of fine-tuning the language models. The generated results are presented and discussed, and a selection of fine-tuned models (and consequently datasets) is made for further evaluation with users based on the quality of the generated descriptions.

4.1 Methodology

To generate descriptions for quests, a selection of models and quests was needed. The selection of models was based on the availability of models suitable to the computational limitations. The quests used to make comparisons should be representative of the entire dataset. These quests were excluded from the dataset before it was used to fine-tune the model. This was done to make sure the quest was unknown to the model for an unbiased generation and evaluation. After this selection was made, the model was fine-tuned on these dataset with the evaluation quests excluded. Three descriptions were generated per quest, and the best description was picked for comparison with the original description.

No direct comparison to previous research could be made to estimate the influence of the inclusion of contextual information during training and generation (RQ2) and the influence of dataset size (RQ3) on generation results. To answer these questions, the researcher made additional comparisons using models trained on datasets without this meta-information and models trained on a smaller subset of the original datasets to estimate the influence of the presence of meta-information and the dataset size on the generation results respectively. Fine-tuning a model on the World of Warcraft dataset without any meta-information also allows for a better comparison between the model used in this research and in the research done by Van Stegeren & Myśliwiec [8], as this dataset uses the exact same structure and only differs in size.

Model selection

The largest computational limitation for this research was the total amount of available video memory (VRAM) of the graphics cards available for use. The fine-tuning process was performed in a Jupyter Lab environment using an NVIDIA A10 graphics card with 24GB of VRAM. While a HPC environment with graphics cards with up to 48GB of VRAM was also available, this did not work properly with some of the libraries used. As such, models were selected that could still be fine-tuned in the Jupyter Lab environment.

It was hard to estimate exactly which models would be suitable for the purpose of this research, taking into account the limitations in computational power and storage. Because of this, a few popular models were chosen that were easily available through the Huggingface Transformers [42] library with PyTorch [43]. These models were GPT-Neo-2.7B [44], GPT-J-6B [45], and BLOOM-7b1 [46].

Quest selection

In order to evaluate the fine-tuned model's performance with users, several quest descriptions should be generated for quests that the model has not seen during fine-tuning. As discussed in Chapter 2.1, multiplayer RPGs often contain relatively straightforward objectives. However, it is still important that multiple different objectives and contexts are represented during the evaluation of the model generation. If every quest objective used in this evaluation set would contain a slightly different version of "Kill this amount of that type of enemy", the results from these quests are likely not representative for the entire dataset. The same goes for quests that are only focused on a single (sub)faction, or quests that all take place in the same game zone. As such, four quests were chosen from each dataset with various differences in objective type, objective complexity, and meta-information such as faction or zone. This way, multiple different quest structures and contexts would be represented in each test set. Objectives can range from traveling to a certain location to talk to a certain non-player character, delivering an item, killing a single enemy or a group of enemies, or finding certain items in the game world. The selection procedure for these quests was done semi-randomly: the researcher used a random number generator to select a quest and then checked whether or not a similar objective or context was already represented in the test set or not. These quests were then excluded from the training dataset so they would be unknown to the language model. If these excluded test quests were part of a quest line, this quest line property was specifically not altered or removed, so that it could be observed whether or not the language model might know to use the context from the other quests in the quest line to set up the correct context for the generated description. After fine-tuning, these test quests were used as input for generation using the language model. This was done with these quests' original objectives and all of their meta-information but without a description as input, so that the language model could attempt to add a fitting description.

Parameters

There are many parameters that can influence the fine-tuning or generation process. These parameters determine or adapt the techniques used during training and generation.

The AdamW optimizer [21] includes various options for fine-tuning, including the use of weight decay and a linearly increasing (or "warm-up") learning rate. The number of warm-up steps for the learning rate was always set to 10% of the total amount of entries in the dataset, after which the full learning rate of 1e-05 was used. This meant 600 warm-up steps for the first two larger datasets, and 120-200 respectively for the smaller datasets. The weight decay was set to 0.1 because of the overfitting observed in the early fine-tuning results. PyTorch also allowed the use of the bfloat16 number format [47] during fine-tuning for reduced memory use. However, the strict memory limitations for fine-tuning the larger models still forced the choice of batch size to be 1 which is very small. This resulted in a longer training time because an update was made to the model for each entry in the dataset.

In the case of GPT-J, the model itself is already about 24GB in size when using 32-bit floating point numbers. Attempting to fine-tune this without any optimizations would require at least 96GB of VRAM and a sizable amount of RAM alongside it. With the optimizations mentioned above, this was reduced to require only about 22GB of VRAM. This does however mean a larger amount of system RAM is required, in this case about 100GB. Since this was available in the Jupyter Lab environment, this did not pose an issue. With this memory restriction out of the way, the largest limitation of this implementation was now time: fine-tuning a single epoch of GPT-J on the larger datasets took about 30 hours each. The code for this project was adapted from DeepSchneider's GPT-Neo fine-tuning example.[10]

[10] Github - DeepSchneider/gpt-neo-fine-tuning-example

All models were trained with similar parameters with some small differences based on observations on the generated results. While some efforts were made to evaluate all of the different parameter options for training and generation as discussed in Chapter 2.3 beforehand, it is hard to evaluate exactly how much influence any particular option had on the generated output. Combined with the fact that it was not clear from the documentation which of these options were mutually exclusive or only available for specific generation strategies, selecting the best options for generating the descriptions was not deemed possible at the time.

4.2 Results

GPT-J seemed to be the most consistent in quality in early fine-tuning experiments, and was also the largest model that could still be loaded and fine-tuned properly with the current limits in computational power. The generation results from GPT-J were considerably more coherent than the outputs from GPT-Neo. While the BLOOM-7b1 model is slightly bigger and newer, it would also run into out-of-memory errors much more often and was therefore deemed unsuitable. Because of this, GPT-J was selected for further use in the experiment. It was also observed that the VRAM requirements differed greatly per dataset. This was likely because the longest input of each training set was used as the maximum input size, and the other inputs were padded to match this size.

In the end, the generation strategy using beam search seemed to give the best results and was therefore selected for further use. Other training and generation parameters such as temperature, repetition penalty, and weight decay differed per model and dataset. Values for all parameters used can be found in Table 6.

Table 6: Training and generating parameters

Game / **Parameter**	World of Warcraft	The Lord of the Rings Online	Neverwinter	The Elder Scrolls Online
Training				
Learning rate	1e5	1e5	1e5	1e5
Batch size	1	1	1	1
Epochs	1	1	1	1
Warm-up steps	600*	600*	200	120
Weight decay	0.1	0.1	0.1	0.1
Generation				
Temperature	0.8	0.9	0.8	0.8
Length (tokens)	120	200	100	75
Repetition penalty	2.0	2.0	1.2	1.2
Beam amount	5	5	5	5

* Reduced to 90 for the partial dataset training

Early experiments showed a large decrease in training loss at the start of each epoch, after which the training loss would stagnate until the start of the next epoch. While it is hard to estimate exactly what happened here, most likely this meant the model picked up on the bracket structure of the quest datasets and quickly learned to replicate this part successfully, causing the training loss to go down fast. First attempts at fine-tuning had a learning rate that was likely too high, and this seemed to result in catastrophic forgetting [48]. This means the model weights are changed too much during fine-tuning and subsequently, the model forgets previously learned text structures. This lead to dubious outputs such as the following description:

> The war against the forces of the furbolg is not only the only thing we know about; it is also a very well. I have been given to know that if they are given enough they will be given on what. An restarding monk inflicvm proceeding spelling underscore Delay soldiers specify counseling Similar Spike SO Zoomcop foregoingocratic━valsameraShortly angel athleteの Peggy leasing thri Childhood kingdoms CummingsammNearly bisexual inception glean prejudices625 inputrellaibia., Douglas counterfe ComedyRE essential Zen detailsloweriance rip monsters guaranteeing loopholeFF volleyballSuch > modelling Rev Plasma Ross Galactic Einstein precinctomial cor pinned Sailor"(playing ambulancemal comprehend cause

When the learning rate was decreased from the initial learning rate of 5e-5 to 1e-5, this behavior seemed to go away. A variable learning rate would likely work best for this to adapt to the speed of training and the decrease of the training loss to prevent the gradient descent from hanging at a sub-optimal local minimum, but evaluating all of the different options for this was not feasible for this research. Instead, a linear "warm-up" learning rate was chosen. This means the learning rate slowly increases from 0 to the chosen learning rate during a set first amount of training steps. This is set to 10% of the total amount of training steps. The number of warm-up steps for the World of Warcraft and The Lord of the Rings Online models was reduced to 90 when training on a smaller subset of 1000 random entries.

4.2.1 World of Warcraft

Quest 1:

Table 7: World of Warcraft - Quest 1

Property	**Value**
Title	What Strange Creatures...
Level	69
Race	Horde
Category	Shadowmoon Valley
Quest giver	Researcher Tiorus
Quest completer	Researcher Tiorus
Quest line	What Strange Creatures...
Required quests	
Unlocked quests	
Objective	Researcher Tiorus at Shadowmoon Village in Shadowmoon Valley wants you to recover 8 Felfire Spleens.

This quest was selected as it is a straightforward kill quest as well as being the first quest in a quest line. This quest's original description is as follows:

> I am interested in learning more about the strange reptiles that mill about the lava flows of Shadowmoon Valley. Never have I seen beasts able to withstand the searing heat of molten lava! Outside of Shadowmoon Village, where the ground has ruptured and lava flows through the cracks, you will find felfire diemetradons. Slay them and bring me their spleens. From their spleens I will be able to begin preliminary analysis of their resistance to heat. Begin your search to the southeast.

Full Dataset

The generated description when trained on the full dataset is as follows:

> It is not too late, <name>. I am still willing to risk it all on a chance that this experiment will succeed. It's hard to believe that these creatures would ever return to their natural state. Now what were we talking about? Oh yes, the spleens of the felfire flayers. If you are willing to go back into the jungle and get me some spleens, I'll reward you with my next experiment. Hurry now!

The generated description is shorter and gives less information about the context of the quest. However, the fact that an experiment is going on for which the researcher needs something and the fact that there will be another experiment may pique the interest of the player. There are also some factual issues, namely that Shadowmoon Valley is actually a barren wasteland and not a jungle as the generated description mentions. The quest is limited in the amount of information it gives the player in the objective, for instance, it does not name the creatures that the spleens should be collected from. As such, the model hallucinates the name of these creatures to be "felfire flayers".

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> You know, I'm beginning to think that there are more strange creatures on this world than I had originally thought. Take for example, the felfire beasts. I've heard that they have some very potent magical properties. You look like you're well versed in these types of things. Why don't you head back out there and try and find me some spleens? I'd be sure to appreciate any help you could provide.

Although the description is close to the original description regarding the magical properties of the animals, this is likely hallucinated since it is not part of the prompt and there is no context in the dataset to link this quest to any other quests in the quest line. There is less mystery and personality in this compared to the full-context version, and not much information on the motivation behind the objective is given.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> I have been working on a new project for the Alliance, something that will help us understand the strange creatures that inhabit this region. I need some fresh samples of felfire spleen and am willing to pay for them. The beasts are quite aggressive, so be careful!

The model generates a motivation for the quest, namely "a new project for the Alliance", although this is still vague. This is also odd, since this quest is limited to players of the opposing faction, namely the "Horde" which is given in the *Race* property. This could mean that the model does not use the *Race* property correctly during generation or training. The model hallucinates the fact that "the beasts are quite aggressive", but since this is a kill quest this is not out of place. Other than that, the description does not contain any hallucinations.

Quest 2:

Table 8: World of Warcraft - Quest 2

Property	**Value**
Title	The Apple Falls
Level	10
Race	Alliance
Category	Rogue
Quest giver	Jannok Breezesong
Quest completer	Syurna
Quest line	
Required quests	
Unlocked quests	Destiny Calls
Objective	Take Jannok's Rose to Syurna in Darnassus.

This quest was selected as it is a travel quest, where players only have to travel to another non-player character and interact with them. Incidentally, it is only available to players with the Rogue class, and has quite a long and complex original description centered around the relation between two non-player characters. This quest's original description is as follows:

> They say I'm 'lovestruck,' whatever that means. Sure, I spend my every waking moment thinking about Syurna. Sure, my home is full of paintings and drawings of Syurna. Sure, I often go days without eating, sleeping, or drinking, while I lament about the love we could have had. Is that so wrong? Now she won't see me! ME! I've tried sending her messages but she won't even talk to another person unless they are a rogue. Could you deliver this flower to her? Don't forget to tell her it's from Jannok.

Full Dataset

The generated description when trained on the full dataset is as follows:

> It is said that a single rose was planted by the Cenarion druids at the falls. The plant is still there, growing and thriving. Take it to my friend, Syurna, and she will help you make use of it. She dwells in the Temple of the Moon in Darnassus.

The generated description turns the quest from a travel quest into a quest that requires fetching an item before travelling to another non-player character. This may be a limitation of the dataset, as some objectives that mention an item to be transported may already give the item in question to the player, while other quests require players to fetch an item and then return it to the quest giver or bring it to another non-player character. This difference is not always made clear in the objective, and as such, the generated output will likely pick randomly. It also refocuses the narrative to be more generic, where the original is clearly meant to be more humorous and focused on the relationship between these two non-player characters. The generated description mentions making use of the rose, but this additional context is also hallucinated by the model. While "the Temple of the Moon" is a real location in Darnassus, it is not given in the prompt and is also hallucinated by the model.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> It is true that you are not a warrior of the Kirin Tor, but we can use all the help we can get. We need more information about this attack, and there is an apple orchard nearby. If you take the rose with you, I will show you where it is. Take it to Syurna at the temple of the Light in Darnassus. She will want to know what you have found.

The model hallucinates a lot of context for this quest, namely the relation to the real faction of the "Kirin Tor", the hallucinated locations "the temple of the Light" and "an apple orchard", and a hallucinated "attack" that supposedly inspires the objective.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> It is time for me to go, <name>. I will leave you with this rose. It is a symbol of our friendship. When you are ready, take it to the druid Syurna at the temple. She will know what to do. Take care, my friend.

The generated description is quite generic, and the quest motivation does not become clearer. While this could be a purposely mysterious setup for the next quest, it is likely not very interesting for the player.

Quest 3:

Table 9: World of Warcraft - Quest 3

Property	**Value**
Title	Raptor Horns
Level	18
Race	All
Category	The Barrens
Quest giver	Mebok Mizzyrix
Quest completer	Mebok Mizzyrix
Quest line	
Required quests	
Unlocked quests	Smart Drinks
Objective	Gather 5 Intact Raptor Horns from Sunscale Scytheclaws, and bring them to Mebok Mizzyrix in Ratchet.

This quest was selected as it is a kill quest that could also be interpreted as a gathering quest because of its wording. This depends on whether or not the model interprets "Intact Raptor Horns" correctly and if it can derive what "Sunscale Scytheclaws" are from this context as this name is not present anywhere in the training set. It also unlocks another quest in which the quest giver aims to make these "Smart Drinks". This follow-up quest is in the training set, and might provide context for this description if the model "remembers" it well enough.

This quest's original description is as follows:

> The raptors of the Barrens are smarter than raptors in other lands. And I think all that smarts is hiding in their horns! If so, then I could grind their horns into powder and use it to make "smart drinks." I could sell them for a fortune!!! And you can help me. Find me intact raptor horns from sunscale scytheclaws. They roam in the southern Barrens, and in the northern Barrens near the border of Ashenvale Forest.

Full Dataset

The generated description when trained on the full dataset is as follows:

> Ain't it a thing how you always be needing a good horn on the back of your belt. You can tell just by lookin' at ya! The horns on the Raptors out in the Barrens are great... big ones. And if you know where to look for 'em, they be close by. So get going now and gather up those horns. Bring 'em to me here in Ratchet. I'll teach ya more about what we got here, and maybe even some of the finer points of our engineering. Now get goin'!

The generated description of this quest is less clear than the original quest, especially because it would likely no longer be correct when placed back into the context of the follow-up quest. The "Smart Drinks" context is completely omitted and instead, the model hallucinates into talking about a "good horn on the back of your belt". If the next quest in the quest line could take this mismatch into account and allow the quest-giver to explain why they wanted these horns in the first place after the player has proven themselves, this may not necessarily be an issue. However, this cannot be assumed. The accent of the goblin quest-giver does come forward prominently. This could be because this quest giver can give quite a few other quests as well and is therefore present multiple times in the training set. The language model also seems to correctly pick up on the fact that the horns need to be found on (living) raptors.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> You have shown that you are willing to help us. But first, I want you to do something for me. I am looking for a special item to make the potions with. The horns of the raptors can be used as the ingredient. If you gather enough, then bring them to me. You will find raptors throughout the lands near the mountains to the east and north.

This generated description is quite close to the original, even mentioning the potions that the quest giver wants to make. This could be because the follow-up quest "Smart Drinks" is still in the training dataset and, although these quests are not directly linked anymore through the *Unlocked quests* property, the model might have made this relation using the quest giver name which is present in both of the quests' objectives. The fact that the player has helped the quest giver before as well as the cardinal directions of the location of the raptors are hallucinated by the model.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> I've been working on a new drink for the centaur. It's called "Sunscale Smarties" and it's made with real raptor horns! I'm not sure how many of these I'll need, so if you can help out, then go ahead and gather some from the sunscale scytheclaws in the Barrens. They're just west of here.

The model somewhat correctly uses the information given in the title and the *Location* and *Unlocked quest* properties. It incorrectly hallucinates a "centaur" character, and the generated cardinal direction is also wrong. The capitalization of the "sunscale scytheclaws" is also inconsistent with the objective.

Quest 4:

Table 10: World of Warcraft - Quest 4

Property	**Value**
Title	It Takes Guts....
Level	75
Race	All
Category	Grizzly Hills
Quest giver	Kraz
Quest completer	Kraz
Quest line	It Takes Guts....
Required quests	
Unlocked quests	
Objective	Kraz at Harkor's Camp wants you to enter Drakil'jin Ruins and collect 5 Drakkari Canopic Jars.

This quest was selected as it is a gathering quest. This quest's original description is as follows:

> Greetings, <class>. It is kind of you to stop and listen to Kraz's tale. For many years Kraz has strained to hear the spirits and do their bidding. Recently, many voices have been crying to Kraz from these ruins. They seek for a peace they cannot find on their own. If you would help the spirits of those who have come before, enter this crypt and collect something that once belonged to those tormented souls.

Full Dataset

The generated description when trained on the full dataset is as follows:

> There is an ancient drakkari shrine to the west, in Drakil'jin Ruins. We used to visit it with our families on long trips. I remember that it was once home to many priceless artifacts. But those days are long past. Now we fear for what happened to the last of our children. If you wish to help me, go to Drakil'jin Ruins and return to me with the canopic jars from the Drakkari dead.

The model hallucinates some odd context for this quest, such as the quest giver visiting this ancient ruin with his family "on long trips" and the concern about "the last of our children". It also rewrites the quest from the third person to the first person, although the model could not know about this beforehand as this is the only quest from this quest giver. The original description does not explicitly name the canopic jars, while this generated one does. This could be seen as an improvement in clarity.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> We must learn as much as we can about this new enemy before it strikes us with its full force. Kraz will tell you more, but for now I want you to head north to Drakil'jin Ruins and bring back any canopic jars you may find. If there are no canopic jars in the ruins, then you might have to kill a few drakkari for me. You'll find the ruins west of here.

The model hallucinates a "new enemy" and incorrectly uses the quest giver name as though they were another non-player character. The fact that the jars can be received from killing enemies is interesting but also incorrectly hallucinated by the model. The cardinal direction of the ruins is also hallucinated.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> You've proven yourself more than a little resourceful, so I'm going to give you a chance to prove your worth. You're going to have to go back into Drakil'jin ruins and get me some jars of the drakkari canopic fluid. The last time I went in there I was almost killed by the spirits. They were very angry about it. Go back in there and get me some jars. If you do, I'll reward you. If you don't, then we'll see how angry they get!

The model generates a logical and coherent motivation for why this quest giver needs help from the player. Similar to the original description, the description involves restless spirits, although in this case the goal is to reduce these spirits' anger instead of bringing them peace. However, the model incorrectly assumes the player has been to the ruins before, which is incorrect.

4.2.2 The Lord of the Rings Online

Quest 1:

Table 11: The Lord of the Rings Online - Quest 1

Property	**Value**
Title	Águti's Advice
Level	54
Category	Moria: The Water-works
Quest giver	Simbi
Quest completer	Águti
Previous quest	
Next quest	
Objective	Simbi is interested to hear from one of his friends, Aguti, about the progress of the expedition in the Water-works.

This quest was selected as it is a travel quest. This quest's original description is as follows:

> Do you have time enough to aid the expedition again? I would like you to speak with a friend of mine, Águti. He has travelled deeper into Moria with the expedition, and I have yet to hear back from him. You can find Águti down in the old Water-works in a place we call the Rotting Cellar, south-west of here, past the Silvertine Lodes. I'm given to understand that it isn't considered the most comfortable post...

Full Dataset

The generated description when trained on the full dataset is as follows:

> Hullo, <name>. I heard that you were looking for something. What can I help you? I'm Simbi, but you may call me Slóthi. It means nothing to me. We've got some problems down here and all we know is that it'll take us days to fix them. I want to make sure my friend Águti is keeping well. He went on ahead with a group of explorers. I haven't heard anything from him since. I need you to look for him at The Deep Descent, north-west of here, if you'd like to help.

The generated description reduces the amount of guidance given by the quest text to find the subject of the quest, although there is more reasoning for why the quest giver can not perform the quest himself. The model also hallucinates into giving the quest giver two names, as Slóthi is actually another non-player character that is not relevant to this quest. The factual information on the location and cardinal direction of the quest completer is also wrong. The correct use of the greeting "Hullo" is positive, as this is very specific to this dataset and shows an adaptation to this writing style.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> Hello there, <name>. You're here to help out? Well, you have. Simbi sent you, and that's always a good thing. He wanted me to ask you to speak with Aguti about our journey into the Water-works. The dwarves at Echad Dúnann have received some complaints about Aguti not returning, and they think he may be in trouble. Speak with Aguti, if you will, and tell him all I've told you!

The model uses the quest giver name incorrectly as if they are another non-player character, as well as hallucinating "complaints" from "dwarves at Echad Dúnann". The wording about the player character having helped out before is also odd. The use of the location "the Water-works" is correctly used from the objective.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> My friend, I have a favour to ask of you. You are familiar with the Water-works, are you not? There is an old man who has been there since before the war, and he may know more of what is happening there than we do. He is called Águti, and he is a great scholar. If you can speak with him, I would be most grateful.

The model hallucinates the fact that the quest completer is "an old man who has been there since before the war" and "a great scholar". It is also not clear why the player is sent to them as no additional explanation is given on "what is happening there".

Quest 2:

Table 12: The Lord of the Rings Online - Quest 2

Property	**Value**
Title	Clear a Path
Level	125
Category	Imlad Morgul: Circle of Wrath
Quest giver	Damrod
Quest completer	Dorgam
Previous quest	
Next quest	
Objective	Orcs have fortified the outer ring of Minas Morgul, to the east. Thin the ranks of the Orcs in this area, so the Rangers can proceed further into the Dead City.

This quest was selected as it is a kill quest. This quest's original description is as follows:

> Thank you for your offer of aid in this dark time; we have greater need of it than ever! Two Orc armies have begun to fortify their positions within the outer tier of Minas Morgul, putting a stop to our forward advance. I sent two Rangers, Dorgam and Dolenthol, westward into the city. They will attempt to gain more information about the area, but could still use your assistance to defeat any of the Orcs you might encounter. Please reduce their numbers as we seek to learn more about the dangers that await us within the city.

Full Dataset

The generated description when trained on the full dataset is as follows:

> You have proven to be a capable and willing Ranger, <name>. We are very fortunate that you were here when we needed help. Now let us continue on with our assault! I am certain that Gothmog's forces are already on the alert. The outer ring of Minas Morgul is currently being occupied by Orcs. These Orcs will hinder our advance as we enter the Circle of Wrath. Head out there and defeat the Orcs encamped there. When you have completed your task, return to me!

This quest rewords the quest to not be about helping two non-player characters gain intelligence on an area, but instead focuses more on clearing the way for a military advance and introduces an enemy called Gothmog, who is an actual named enemy encountered in the game and is named in several other quests in this quest category. This can be considered positive as the model correctly uses additional context available in the training set. The model correctly names the context of the "Circle of Wrath" that was provided in the *Category* property, and uses the "outer ring of Minas Morgul" provided in the objective correctly as well. The model fails to use the *Quest completer* property correctly here, as the generated description instructs the player to return to the quest giver when they are done, which is incorrect.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> You have proven to be a capable and willing Ranger, <name>. You should be ready to move on from here. If you wish to do so, I suggest that you make your way back to Echad Eregion and give it my good wishes. I am certain that Elrond will appreciate your aid. He may have already told you what has been planned.

It is interesting to see the exact same use of the starting line as the generated description with full context, since this particular phrase is not used anywhere else in the training dataset. The location "Echad Eregion" is hallucinated by the model, as well as the involvement of a character called "Elrond". The wording of "giving good wishes" to a location is also odd.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> You are here to help us? You will need to prove yourself worthy of that honour, my friend. I am Damrod, leader of the Rangers of Imlad Morgul. Our numbers are few, but we fight with great valour. We must thin the ranks of the Orcs in these parts, so that we may advance further into the city. Will you do this for us, and return to me when you have completed your task?

The model correctly uses the *Quest giver* and *Category* properties, but not the *Quest completer* property since the player is incorrectly instructed to return to the quest giver afterwards. The model incorrectly assumes the quest giver is the leader of these Rangers, and that the groups of Rangers is called "Rangers of Imlad Morgul" which is incorrect.

Quest 3:

Table 13: The Lord of the Rings Online - Quest 3

Property	**Value**
Title	From Whence They Came
Level	74
Category	Gap of Rohan: Heathfells
Quest giver	Orc-scimitar
Quest completer	Ondor
Previous quest	
Next quest	
Objective	While fighting Orcs in the Heathfells, you found a sword one of them dropped. While it bore the usual crude hallmarks of Orc-crafting, it was very nearly brand new, suggesting that Orcs must have a lair nearby where they are forging these weapons.

This quest was selected as it is a travel quest initiated by an item looted by the player. The objective is relatively long so this might be a challenge for the model.

This quest's original description is as follows:

> This blade was taken from a fallen Orc in the Heathfells. While forged in the usual crude and utilitarian manner of the Orcs, it appears to be almost brand new. Somewhere nearby must be a cave or encampment where these particular Orcs are forging weapons. If you can find it, you should be able to tell the Rohirrim of its presence so that they can take action against them.

Full Dataset

The generated description when trained on the full dataset is as follows:

> This sword bears an unusual craft-mark, but its origin is clear to me. It came from an Orc-horde dwelling nearby. You should take this weapon to Ondor and see if he knows anything about it.

The objective of this quest is relatively long but doesn't give a very clear assignment, which likely confuses the model. Instead of focusing on finding the orc lair mentioned in the objective and finding the quest ender there, it instead focuses the quest completely on bringing the found weapon to the quest ender Ondor. This may be a limitation of the dataset, as some quests have multiple objectives contained within a single quest. In this case, the player is supposed to search for an orc camp, after which they receive a new objective to talk to Ondor there. Since only one objective is given for each entry in the training set, the model cannot know that these multiple steps could be involved. It does correctly use the fact that the Orc lair is nearby and even calls it an "Orc-horde dwelling" which is interesting. Calling groups of Orcs "orc-horde" is done several times in the training set, showing that the model adapts to the word use in the training set and understanding this word in context. It is the only generated description from the Lord of the Rings Online test set that does not use the name tag, but this makes sense as there is no character talking to the player.

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> You found this sword as an Orc fell from his horse while attempting to flee. You should take it to Isengard and tell Saruman what you learned. He may know something about the sword or its origin.

The involvement of an Orc falling from his horse is hallucinated by the model, as well as the involvement of a character called "Saruman" and a location called "Isengard". Similar to the generation result from the model trained on all of the context, the description is relatively short.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> This sword bears the usual hallmarks of Orc-crafting, but it is also nearly brand new. It suggests that there is a lair nearby where Orcs are forging these weapons. We should investigate and see if we can learn more about this threat.

This description does not contain any hallucinations as it rephrases the objective almost completely.

Quest 4:

Table 14: The Lord of the Rings Online - Quest 4

Property	**Value**
Title	Packing for the March
Level	105
Category	Pelennor (After Battle)
Quest giver	Angbor
Quest completer	Angbor
Previous quest	Testing the Defences
Next quest	
Objective	Angbor has been tasked with preparing for the march eastward and must gather supplies for the journey.

This quest was selected as it is a gathering quest. This quest's original description is as follows:

> We are preparing to march soon. We must start packing the supplies that we will need on this journey to the Black Gate. Collect supplies from around the pavilion, then pack them in crates on one of the carts.

Full Dataset

The generated description when trained on the full dataset is as follows:

> We have spoken about our plan to leave, <name>. The Orcs will be there, and I would like to prepare ourselves. I need you to make your way through the rubble here in Pelennor and find some supplies that we might bring along. There are a few things that we will need, and each of them can be found around this place. We will not be leaving until we have all of these things.

The generated description is longer and more detailed than the original, and focuses more on the urgency of the task. There is some questionable grammar, such as "the orcs will be there" instead of something along the lines of "the orcs will be here soon". The same goes for "I would like to prepare ourselves", where something like "I would like us to prepare ourselves" may be more conventional. This could be a consequence of the dataset as the old-fashioned Tolkienesque writing style is often used, and this may confuse the model. The description also assumes the player has previously spoken with this non-player character about their plan to leave. This could be because of the previous quest given in the meta-information, but this is still an assumption by the model. The model correctly uses the location context provided by the *Category* property by naming the location "Pelennor".

Removed Embedded Meta-information

The generated description when trained on a version of the dataset with only the quest titles, objectives and descriptions present is as follows:

> Hail, <name>. You have proven a great aid to Angbor's army in recent days. We have much work to do and we cannot waste any more time! The first step is to make sure we are ready. There are many who have volunteered to assist us but there are also many who have not. Those who have not should be informed

> that we will need their assistance if they choose not to come. They may think that they will not see anything of note during this long march, but it would be better to prepare than to be unprepared. Speak with Núrthauk, an Elf here at Gath Forthnír, and ask him what needs doing. He will know where to direct you.

The use of "Angbor's army" is wrong as Angbor is the quest giver. The involvement of another character "Núrthauk" and a location "Gath Forthnír" is hallucinated by the model. There is also some weird wording throughout the description, and the motivation behind the objective does not become clearer.

Partial Dataset

The generated description when trained on a random subset of 1000 entries from the original dataset is as follows:

> You have done much to aid us, <name>. You will be a great asset to us in the days ahead. I would like to ask you to help us by gathering supplies for the march. There is much to do before we depart. We will begin at once.

The model makes some assumptions about the player's previous involvement, this may come from the "(After Battle)" information in the *Category* property. However, there are no hallucinations in this generated description. The description is straightforward, and no motivation for the objective is given.

4.2.3 Neverwinter

Since this is already a relatively small dataset which does not contain much contextual information, it is assumed that fine-tuning an LLM on an even smaller subset of this data or on the same dataset without meta-information will likely not provide additional insights into answering RQ2 and RQ3, and consequently, these steps were not performed for this dataset.

Quest 1:

Table 15: Neverwinter - Quest 1

Property	**Value**
Title	Thunderhead Reach: Target - Grackel
Level	69
Quest giver	Pirri the Swift
Quest giver location	Spinward Rise
Quest completer	Pirri the Swift
Quest completer location	Spinward Rise
Previous quest	
Next quest	
Objective	Kill the owlbear Grackel at Thunderhead Reach in Spinward Rise.

This quest was selected as it is a kill quest. It focuses on the extermination of a single named monster.

This quest's original description is as follows:

> Some beasts are too smart for their own good. A vicious owlbear named Grackel has been trained by the Cult of the Howling Hatred to sniff out our spies. This has to end! Head to Thunderhead Reach and take out Grackel.

This quest's generated description is as follows:

> A owlbear named Grackel has been a thorn in our side for as long as Spinward Rise has been in existence. He's been responsible for the deaths of many, including our own soldiers. We need you to take Grackel out. Once and for all.

The generated description correctly uses the name Grackel, and the fact that it is an owlbear. The model hallucinates that this monster has been around since the inception of the location Spinward Rise, and that it is responsible for many deaths, as this is not known from the prompt. The original description mentions the monster being trained by the cult, but since this is not present anywhere else in the prompt, the model does not know about this connection.

Quest 2:

Table 16: Neverwinter - Quest 2

Property	**Value**
Title	Warding off Madness
Level	20
Quest giver	Councilman Graves
Quest giver location	Reclamation Rock
Quest completer	Councilman Graves
Quest completer location	Reclamation Rock
Previous quest	Call to Alms, Recovering Reconstruction
Next quest	The Verge of Collapse
Objective	Gather Materials around Helm's Hold to Reinforce the Wards on Hexabaal.

This quest was selected as it is a gathering quest that is part of a quest line. This quest's original description is as follows:

> There is an Aberrant creature called Hexabaal still trapped in our world, but the barriers keeping her contained are beginning to weaken. The last thing we need is an aberration running loose alongside these Black Earth cultists. We need to strengthen the barrier keeping Hexabaal in place. I've got a list here compiled by Hedgewizard Nottin that details what is needed to reinforce the barrier. Can you handle this before Hexabaal is loosed upon this plane?

This quest's generated description is as follows:

> We've got a real problem here. The Cult of the Black Earth has been attacking Helm's Hold with magic and madness. Our wards are weak and our defenses are low. If you can find some magical materials to reinforce our wards, we can keep the cultists from breaking through.

The generated description is quite short and misses a lot of context given by the original description, but the model still correctly uses the context of the cultists. However, it hallucinates the cultists to be "attacking [...] with magic and madness" which cannot be assumed from the prompt. This is probably taken from the title, but without the context of this creature, it becomes incorrect.

Quest 3:

Table 17: Neverwinter - Quest 3

Property	**Value**
Title	A Sign of Good Faith
Level	60
Quest giver	Drann Barrow
Quest giver location	Whispering Caverns
Quest completer	Drann Barrow
Quest completer location	Whispering Caverns
Previous quest	An Unlikely Alliance
Next quest	Paper Trails, Ritual Behavior
Objective	Destroy the Drow Poison Shipments in Zasraena.

This quest was selected as it is a travel and interact quest. Additionally, the original context for this quest is quite complicated: you are given information on a third party by a relatively new ally as a means of showing companionship in the struggle against a fourth party. This quest's original description is as follows:

> I sense you have your suspicions about working with me. Perhaps this will prove I am sincere. The drow operate a lucrative trade in rare and expensive poisons. We have been known to purchase some of them ourselves for certain... clients. To show I am in earnest, I am prepared to share with you the location of several such shipments. I trust you realize that this jeopardizes an extremely valuable association with the drow. It is not something we would do unless we felt the Cult of the Dragon posed an even greater threat to us all.

This quest's generated description is as follows:

> The drow have been sending the poison to their allies for use as a ritual component. If you were to destroy these poison shipments, it would be a great help to us in our fight against the drow.

The generated description is quite short, and the model hallucinates "a ritual component" and "our fight against the drow" into the backstory which is not fitting.

Quest 4:

Table 18: Neverwinter - Quest 4

Property	**Value**
Title	Assembling the Future
Level	10-70
Quest giver	Sybella Artis
Quest giver location	Protector's Enclave
Quest completer	Sybella Artis
Quest completer location	Protector's Enclave
Previous quest	Collecting the Past
Next quest	Searching the Present
Objective	Build a working Scrying Orb

This quest was selected as it is a quest that requires the player to perform several steps. This quest's original description is as follows:

> Now that we have a supply of raw materials, we must see what can be salvaged and attempt to recreate one of the Order's orbs. This process may take some time, but I am confident we will succeed.

This quest's generated description is as follows:

> I've been studying the Astral Plane. I believe it is possible to scry into the future. I've seen it done on the astral plane, so I'm sure it can be done here too. I have some of my astral energy left, and I'm going to use it to make a working scrying orb. I'll need your help to build it.

The generated description is longer than the original description, and hallucinates quite some additional backstory that does not line up with the storyline of the quest. The model inconsistently uses capitalization for "the Astral Plane", although this may be because both methods of capitalization are used in the dataset. This quest is part of a relatively long questline but the model does not seem to have made a link to these quests using the *Previous quest* and *Next quest* properties during generation. Since the objective is unclear, it may be hard for the model to add a fitting backstory that remains correct.

4.2.4 The Elder Scrolls Online

It can be observed from the following examples that this dataset does not contain a lot of context for the model to generate from, and many of the original objectives do not contain detailed objectives. Although the meta-information seems to help the generation somewhat, much of the context for the quest needs to be thought up by the LLM as seen in quests 1 and 4. If this additional context is not thought up by the LLM, the description will become straightforward as seen in quests 2 and 3. Similar to the Neverwinter dataset, it is assumed that fine-tuning an LLM on an even smaller subset of this data or on the same dataset without meta-information will not provide additional insights into answering RQ2 and RQ3, so these steps were also not performed for this dataset.

Quest 1:

Table 19: The Elder Scrolls Online - Quest 1

Property	**Value**
Title	Lady Laurent's Favor
Level	32
Quest type	Side
Faction	
Quest giver	Stibbons
Location	Alik'r Desert
Previous quest	
Next quest	
Objective	Help the Lady's butler recover her tools.

This quest was selected as it is a X quest. This quest's original description is as follows:

> Lady Laurent is displeased with her butler, Stibbons, for leading the Elves to Salas En. To get back in her good favor, he's asked me to help him by gathering Laurent's tools from among the ruins.

This quest's generated description is as follows:

> Lady Laurent has lost her memory and her tools. Stibbons has asked me to recover her tools and help her regain her memory.

The generated description is very short, and the model seems to be hallucinating that the "Lady" character has memory loss.

Quest 2:

Table 20: The Elder Scrolls Online - Quest 2

Property	**Value**
Title	A Time for Mud and Mushrooms
Level	
Quest type	Guild
Faction	Psijic Order
Quest giver	Loremaster Celarus in the Ceporah Tower's main hall or Josajeh in the Athenaeum of Ceporah Tower
Location	
Previous quest	
Next quest	The Towers' Remains
Objective	Seal time breaches in the swamps of Deshaan and Shadowfen.

This quest was selected as it is a travel quest that makes the player interact with several objects spread across a game region. This quest's original description is as follows:

> My allies in the Psijic Order discovered another cluster of time breaches in the northernmost regions of Black Marsh. I need to seal them.

This quest's generated description is as follows:

> Loremaster Celarus has asked me to seal more time breaches in the swamps of Shadowfen and Deshaan.

The generated description is very straightforward, almost completely repeating the objective. As such, it does not give any additional motivation for the objective. It does use the quest giver property correctly, although only one of the two potential quest givers are mentioned. This could be because the dataset contains two other quests that also concern "time breaches" that mention this quest giver by name. The model also correctly picks up on the fact that this assignment has been given before as it uses "more" without being instructed to. This could be because there are six quests in total in the dataset that mention "time breaches", all of which are given by this quest giver and that also has the same tags for Quest type and Faction.

Quest 3:

Table 21: The Elder Scrolls Online - Quest 3

Property	Value
Title	The Hedoran Estate
Level	
Quest type	Side
Faction	
Quest giver	Prefect Antias
Location	Cyrodiil
Previous quest	
Next quest	
Objective	Investigate a destroyed manor that was overrun by bandits.

This quest was selected as it is a gathering quest. This quest's original description is as follows:

> The Hedorans were killed by bandits. Prefect Antias has asked me to recover anything I find in the remains of their home.

This quest's generated description is as follows:

> Prefect Antias asked me to investigate the manor of the Hedoran Estate. The manor is overrun by bandits and has been destroyed.

This generated description is very short, similar to the previous quest. The model does combine the title, quest giver and objective properties together properly.

Quest 4:

Table 22: The Elder Scrolls Online - Quest 4

Property	**Value**
Title	A Traitor's Luck
Level	
Quest type	Side
Faction	
Quest giver	Note from a Bottle
Location	Reaper's March
Previous quest	
Next quest	
Objective	Help uncover evidence that a Khajiit is selling Dominion secrets to the enemy.

This quest was selected as it is a quest received from an item. This quest's original description is as follows:

> A hidden note in a bottle has led me on a trail into the town of Pa'alat.

This quest's generated description is as follows:

> I found a bottle containing a note from a Dominion spy named Razum-dar. He was captured by the Dominion while trying to recover some important documents.

The model hallucinates an existing character "Razum-dar" that is not directly related to the current quest in the generated description. This character is a "Khajiit" involved with the "Dominion" faction which is a relatively good connection made by the model, but it is still factually wrong. The capture of this character is also hallucinated. The use of "a bottle containing a note" from the quest giver property is good.

4.3 Discussion

Some clear differences in the generated results of these datasets can be observed, as well as differences between the generation results of different levels of context and amount of entries. This confirmed some of the identified issues with these datasets from Chapter 3.5.

Observations

From these results, several elements can be identified that need to be present in both the objective with meta-information (input) and the description (output) for a coherent and logical description that is similar to the original description. The first element that needs to be present is the assignment itself, since the player needs to know what they have to do. This element is present in the dataset as the objective, and is linked more strongly to the description when the objective is repeated there. The second element is the context in which this assignment takes place, as the player needs to know what locations and characters are involved, so they know where they need to go to perform this assignment. This is currently provided in the additional meta-information provided in these datasets, which is linked to use of these terms in the descriptions. The usage of this information in the descriptions differs per dataset and even per entry in each dataset. However, it can be observed that the model generally uses this information correctly when it is present in both the input and the output. The last element that needs to be present is the motivation behind this objective. However, this is only sporadically mentioned in the objectives and is generally not present in the currently available meta-information either. As such, the model learns about the use of different types and styles of motivation but likely fails to link this to any of the input text. The current results show that this motivation is generally the weakest point in the generated descriptions, as this is where the most hallucinations are observed.

Across all datasets, many quest objectives do not contain or describe the motivation for the quest, which is likely the largest reason for the assumptions made during the generation of the description and subsequently, the hallucinated reasoning for the objective is often wrong. As discussed in Chapter 3, the Lord of the Rings Online dataset contains many unclear quest objectives, but the quests with clear objectives often contain clear reasons for those objectives as well, as seen in quests 2 and 4. Having this information available in the objective and expanding on it in the description likely allows the model to more easily link the word use between objectives and descriptions. Quest 1 contains some motivation in the objective, but not as much context is provided compared to quests 2 and 4 and as such, the model needs to make more assumptions when generating this description. Comparatively, quest 3 does not provide any motivation in the objective and contains more hallucinations than the other three quests as a result. The context information will likely be more important in informing the model on generating this motivation in these cases, which is observed when comparing the results from the full datasets to the results from the models trained without meta-information.

Evaluation

While some confusing wording and factual mistakes remain in many of these examples of generated descriptions, the use of additional context was observed to reduce the amount of mistakes in referencing relevant characters and locations. The addition of this meta-information generally improved generation results in these larger datasets. The differences in how the objectives were worded in each dataset and the way this influenced generation results has also become clearer, as the two largest datasets also contained the most descriptive objectives which may have improved their performance.

The results from the World of Warcraft and The Lord of the Rings Online models were quite good, although there were still instances of factual mistakes and hallucinations, or questionable sentence structures. It can be observed that the model often uses the contextual meta-information correctly, especially properties like the names of quest givers, quest completers, and locations. This becomes especially clear in comparisons with the results of models trained on datasets without this meta-information. However, without further analysis into the attention values within the model, it is hard to estimate the exact influence of this additional context. The properties for questlines and previous or next quests were not used consistently and are rarely mentioned directly. As such, it might be hard for the model to directly link certain word usage to the presence of this information, and it may only serve to confuse the model.

Unfortunately, the results of the Neverwinter and Elder Scrolls Online are relatively poor. The short objectives and descriptions of the Neverwinter and The Elder Scrolls Online datasets limited the amount of background information that the model could use as a basis for the description generation. As such, the generated descriptions for these models are generally either uninteresting to read because the model only used the limited amount of available information, or the descriptions contain many factual mistakes because the model made incorrect assumptions about the context. This may be because the objectives in these datasets are used more as overarching guidelines, since there are often multiple steps during these quests that are not currently included alongside the objective. The same is true to some extent for the Lord of the Rings Online dataset, where quests can contain multiple steps which are not included in the quest objective. In contrast, the entries in the World of Warcraft dataset generally contain all of the required knowledge for the player to successfully complete the quest, as the description and objective is generally the only information the player receives. Quests do not have multiple steps, as this will generally be split into multiple quests that form a questline instead. As such, this measure of granularity likely means this is the most suitable dataset for this purpose. This also indicates a need for more complete and granular quest information for accurate generation.

By comparing the results from the models trained on datasets without meta-information to the results from previous research, the differences between each of these models can be observed. Factual mistakes such as references to the quest giver in the third person and the introduction of irrelevant characters or factions can often be observed. These types of errors were not unexpected, since this dataset structure without meta-information is very similar to the structure used in previous research [8]. As such, similar errors were likely to emerge. However, the use of newer models did show an increase in coherence in the generated outputs when directly comparing these results. Comparing the current results with the results from the models trained on the full dataset with context, it becomes clear that the model needs to make more assumptions about the context when it does not have any information on this, which generally reduces the quality of these results.

The models trained on a random subset of 1000 quests performed quite well, sometimes matching or exceeding the full dataset in clarity and lack of hallucinations. However, this also led to a lack of spontaneity as the model had a tendency to copy the objective quite directly without adding much context. Since three descriptions were generated per quest and one of these was picked manually as the final result, an increase in variety of motivations and quality was noticeable during this selection process. The results from the model trained on the full dataset were observed to differ only slightly in wording and background per quest, while this variety was much higher for the smaller dataset. In these results, all three descriptions often contained very different backstories and motivations for the same quest. This may indicate a balance between output variety, and dataset size and complexity.

Selection for further evaluation

From this preliminary evaluation, a selection of datasets was made for further evaluation with users. The generated descriptions from the World of Warcraft and The Lord of the Rings Online models were deemed to be the highest in quality. As such, these results were selected for further evaluation with users.

Additionally, the provided meta-information was observed to increase the quality of generated results in terms of coherence and contextual correctness. As such, the models trained on the full datasets with context were selected for further evaluation with users.

Limitations

Although some attempts were made to generate quests with a variety of parameters and compare these results objectively, it was hard to estimate the influence of each parameter and the resulting quality of the descriptions. As such, the current parameters were an estimation of the best performance, but better performance could still be achieved with these models and datasets through further parameter adjustments. Similarly, three results were generated for every entry in the results, of which the best was selected by the researcher. This may have added a subjective bias from the researcher to the quality of the results.

The random nature of the quest selection revealed some mistakes that were still present in some of the datasets even after data cleaning, such as quests having multiple quest givers in The Elder Scrolls Online and Neverwinter datasets. However, this is hard to prevent when manually sanitizing datasets of this size.

The repetition of the objective in the description is present mostly in the World of Warcraft and Lord of the Rings Online datasets, and much less so in the Neverwinter and The Elder Scrolls Online datasets. This is a likely reason for the reduced quality of generated results for these smaller datasets, alongside the reduced amount of information to inform the model on the context of the objective. Removing quests without a clear motivation from the quest giver may show an increase in generation performance.

Factual mistakes are still present in these results, since the given meta-information only provides a small part of the entire quest context and information. Although the model may be informed on the names of certain game zones or characters, the properties of these named entities are not clearly provided and as such, assumptions need to be made by the model during generation. Additional efforts are needed to determine the actual most relevant context information in these datasets, as removing superfluous information will likely improve the model's generation performance. Although some preliminary filtering of low quality quest texts was already performed, this may need to be extended as the Lord of the Rings Online and World of Warcraft datasets still contain quests without clear objectives and descriptions.

The method of selecting 1000 quests through a random sample was not perfect, as this subset contained a lot of partial contextual information that the model may not be able to use effectively, and might even serve to confuse it. This could for example be references to quests or characters that are no longer mentioned in the database but are still present in some quests' meta-information. Another consideration for this was to use the first 1000 quests from each datasets as most datasets are roughly sorted by required player character progression like character level or story progress. As such, these first 1000 quests would likely be more related to each other. However, this would also have led to a bias towards these relatively early quests.

5 Quantitative evaluation

Hamalainen & Alnajjar [40] describe the need for multiple types of evaluation to accurately assess the limitations of such a creatively oriented text generation system. As such, a digital survey and subsequent in-person group interview with interested survey participants were chosen as the primary methods of evaluation. This chapter will discuss the methodology and results of this digital survey, while Chapter 6 will discuss the focus group interview. The results from the survey can be used alongside the information gained from the focus group interview to answer RQ1.

5.1 Experiment design

A selection of original and generated descriptions from the same quests were independently evaluated by participants so that their results could be compared directly. This was done on several measures of text quality.

Measures

Appropriate focus points for the survey can be set up when the most relevant components of computational creativity are identified and applied to this context of quest generation. Many different components for computational creativity were identified by Jordanous [41] and Hamalainen & Alnajjar [40]. In the case of creative text generation for quest descriptions, from Jordanous' collection these were decided to be Generation of Results, Value, and Originality. These components can also be found among the papers collected by Hamalainen & Alnajjar. Some other relevant components identified in these papers are relatedness, topicality, and coherence.

These relevant components were grouped into several overarching factors relating to textual quality. This first aspect was creative value, which means the description should improve the player's experience. The second aspect was consistency, which means the description needs to be coherent and understandable. The third aspect was contextual fit, as the description needs to make sense within the given context. However, these aspects are still open to interpretation and are not yet suitable to put into a survey as is. Based on these overarching factors, descriptive keywords were ideated. These keywords could have been used as measures to be rated on a Likert scale in their current form, but it was decided to incorporate each keyword into a statement instead since some of these keywords already turned out to be descriptive sentences.

The list of ideated keywords per construct can be found in Table 23. These descriptions were then compared to remove superfluous or ambiguous descriptors of similar concepts, and reduced to the 10 statements found in Table 24. Additionally, because of the fantasy background of the datasets used, these statements were required to be independent of the participant's knowledge as much as possible. This was decided because questions like the correct use of English words or the realism of the scenario are strongly influenced by a participant's prior knowledge of the setting and should preferably be avoided.

Table 23: Ideated keywords

Creativity	**Consistency**	**Fittingness**
Engaging	Correct grammar	Motivation is clear from description
Good	Correct spelling	Objective and description fit well
Original	Clear storyline	Description matches objective
Enjoyable	Easy to read	Clear purpose
Surprising	Flows well	Clarity
Contributes positively	Understandable	Fits the context
Adds value	Makes sense	Topical
Important	Coherent	
Necessary		
Novel		
Unexpected		
Immersive		
Unpredictable		
Motivational		

Table 24: Ideated statements

Construct	**Statement**
Creativity	The description is engaging.
	The description is original.
	The description makes the quest more enjoyable.
	The description is surprising.
Consistency	The grammar in the description is correct.
	The description has a clear storyline.
	The description is easy to read.
Fittingness	The description matches the objective.
	The description fits the context of the quest.
	The purpose of the quest is clear from the description.

Survey Design

The survey was set up using the Qualtrics survey platform, which means it was possible to use more advanced randomization compared to other platforms such as Google Forms or Microsoft Forms. This meant that for each quest there was a version with the original description and a version with a generated description present in the survey. Participants randomly evaluated either the original or generated version of that quest. Since each participant only saw one version of each quest, the survey results for these versions were independent and could be compared directly.

There is a limit to the amount of survey questions before the survey becomes too long for participants. As such, only the two best-performing datasets were selected for the survey. The selection of these datasets was done by the researcher as described in Chapter 4.3. All four quests from both World of Warcraft and The Lord of the Rings Online were evaluated, for a total of eight quests with two versions each. By evaluating multiple quests from multiple games, variance between games and individual quests could be accounted for.

To limit any learning effect between participants, both the order of quests and the order of questions were randomized once at the start of the survey. This way, the question order was different per participant but consistent for each individual participant. Participants evaluate a version of each quest on the 10 statements as seen in Table 24.

Although 5-point Likert scales are the most commonly used as shown by Hamalainen Alnajjar [40], for this survey 7-point Likert scale questions were chosen for their increased granularity over 5-point scales as advised by Van der Lee et al. [39]. This 7-point scale contains the values Strongly Disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, and Strongly Agree.

Recruitment and procedure

Participants were recruited through several channels, namely through several Whatsapp group chats from the study of Interaction Technology and its study association's communities, as well as several of Reddit's community forums dedicated to videogames (such as /r/truegaming) and survey sharing (such as /r/SampleSize), and the Discord server from the university's student esports association. The only requirements for participation was for participants to be at least 18 years old. A personal interest in video games or storytelling was mentioned as being a plus but was not set as a strict requirement.

After opening the survey, participants were asked to give consent on the gathering of their data for the purpose of this survey. This consent form can be found in Appendix A. Next, several demographic questions were posed on the participants' age group, gender, and experience with the four multiplayer RPGs used in this research. Although only quests from two of these games would be present in the survey, the average experience of these users with all of these games might still provide insights into their performance. After this, participants would go through the eight questions of the survey. After these questions, participants were prompted to leave their email address if they were interested in attending a follow-up focus group interview. Additionally, participants were asked whether or not they would like to share any comments on the survey.

5.2 Results

5.2.1 Participants

In total, 78 participants started the survey. Of those, only 36 fully completed the survey. While it is not uncommon to use partial survey responses, in this case only the completed survey responses were used for the statistical analysis to maintain sample size consistency and because the participants of the partial entries were likely not as interested or invested in the survey's topic. As such, their partial entries may not be representative of the overall population. One of the entries was not considered to be serious as the participant only gave ratings of 1 (Strongly Disagree) or 4 (Neutral) and went through the entire survey in only three minutes so their entry was discarded.

Participant demographics can be found in Table 25, while the participants' experience with the games used in this research can be found in Table 26.

Table 25: Participant demographics

Age	Amount	Percentage
Under 18	0	0%
18-24 years old	15	43%
25-34 years old	16	46%
35-44 years old	3	8%
45-54 years old	1	3%
55-64 years old	0	0%
65+ years old	0	0%
Gender		
Male	24	68%
Female	9	26%
Non-binary	0	0%
Prefer to self-describe	1*	3%
Prefer not to say	1	3%

* This person self-described as genderfluid

Table 26: Participant experience

Experience \ Game	World of Warcraft	Lord of the Rings Online	Neverwinter	The Elder Scrolls Online
I have never heard of this game before	0	5	17	1
I have heard of this game but never played it	16	22	12	24
I have played this game in the past	16	8	6	9
I currently play this game	3	0	0	1

5.2.2 Methodology evaluation

Although not necessarily required for a relatively simple survey as used in this research, it is still beneficial to evaluate the research methods used as it allows for stronger conclusions on the gathered survey data. While this is generally an iterative process, this was not possible because of the limited scope of this research. Instead, this evaluation was performed with the intention of providing potential future research insights on whether or not the statements and constructs used here should be improved upon.

Common factor analysis

The aim of this method is to estimate how well the descriptive statements that were previously set up compare on measuring the same construct. As mentioned previously, the constructs defined earlier are hard to measure directly. By performing a confirmatory factor analysis, the validity of these constructs can be estimated by measuring the strength of the relationship between each construct's statements. This is based on the assumption that all of these factors describe their construct similarly and equally, and that this can be described by a formula. The sample data is then compared to this hypothesized model, and the differences found there can be used to estimate the quality of the statements describing their construct. This is done with several different fit indicators, each focusing on different statistical properties to describe different aspects of the quality of the model. Some examples of properties that are taken into account in this are the variances (the amount of deviation from the mean in a single random variable), the covariances (the similarity of variances of multiple random variables), and residuals (the difference between data from the sample and the estimated data from the model).

The standardized weights per question, covariances between constructs, and residuals of questions can be seen in Figure 3. In this graph, the bottom squares are the survey questions (x1 being question 1, x2 being question 2, etc.), with the values below them indicating the residual error, and the arrows from the constructs above them indicating their standardized weights. Preferably, the standardized weight of each question is high while the residual error is low. This would indicate the question represents its construct appropriately and there is not a lot of unexplained spread in the question scores. The covariances between the three constructs at the top of the graph describe how closely the construct scores move together. As mentioned previously, since all three constructs are related to the textual quality of the descriptions, it is not surprising that the covariances between them are also relatively high. It can be observed that questions 1 and 4 are of quite high quality, while questions 3 and 5 are quite poor in relation to the other questions in their respective constructs.

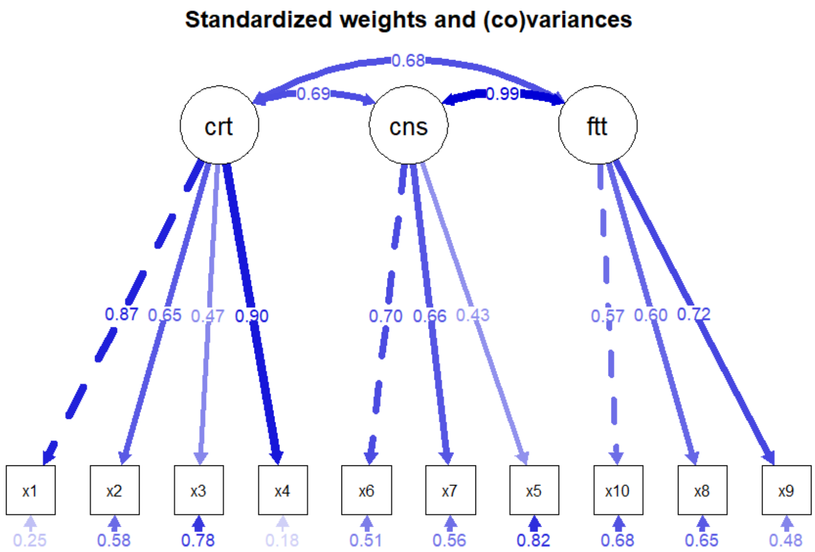

Figure 3: Overview of loadings on constructs of Creativity (crt), Consistency (cns) and Fittingness (ftt), with covariances between the constructs at the top and residuals per question at the bottom

The confirmatory factor analysis showed that all questions load (i.e. contribute to the respective construct) significantly on each construct ($p < 0.001$) and are therefore related to each other enough to properly form a construct. Similarly, the constructs are correlated with each other significantly ($p < 0.001$) which indicates they all describe a larger concept, which in this case is textual quality.

Model fit indicators

Hooper et al. [49] describe several indicators that can be used to estimate how well the hypothesized model (in this case, the constructs split into their respective statements) fits with the actual sample data that was collected in the survey. Many different indicators can be used to estimate this model fit based on sample size and . Hooper et al. [49] recommend reporting at least the Chi-square test, the RMSEA, the SRMR, the CFI, and the PNFI. The values of these indicators as well as a short description of each indicator are as follows:

The Chi-square indicator estimates the size of the difference between estimated and actual covariance values. Its p-value equals 0.00 ($t = 203.74$, $df = 32$) while it should preferably be greater than 0.05. This indicates a poor fit. However, it is quite sensitive to small sample sizes so it is reported here for completeness but should not be a guiding factor in this case.

The root mean square error of approximation (RMSEA) is a test on how well the parameters would fit with the estimated population covariances if these parameters would be optimally chosen. In essence, it reports how far the model is from being a perfect model. This value equals 0.138 (90% CI [0.121, 0.157]) while it should preferably be smaller than 0.07. This indicates a poor fit.

The Comparative Fit Index (CFI) indicates the difference between the model and a null model where no factors are related. This indicator works better for small sample sizes. Its value can range from 0 to 1, with higher values indicating a better model. In this case, its value equals 0.847 while it should preferably be greater than 0.95. This indicates a somewhat poor fit.

The Standardized Root Mean Square Residual (SRMR) is an estimate of the difference between the sample covariance residuals and the covariance residuals of the hypothesized model. Its value equals 0.086 while it should be smaller than 0.08 and preferably even smaller than 0.05. This indicates a somewhat poor fit.

The Parsimonious Normed Fit Index (PNFI) indicates how close the model is to the observed covariances from the sample data, favoring simpler models over more complex models. Its value is 0.587 and it should be greater than 0.50. This indicates an acceptable fit.

Unfortunately, the indicators for a good fit are not as desirable as hoped for this model. In this case, all but one of these indicators were off from an adequate fit. This does not mean this model is necessarily bad, but that the reality behind this dataset is more complex than this model is currently set up to represent. For instance, a question regarding grammar may load on both consistency and fittingness, or there may be more factors than the ones that were ideated that would describe the construct more completely, or certain factors are too similar and measure the same aspect of a construct too closely. These points would not be taken into account in the current version of the model, and could therefore lead to a bad fit of the currently proposed model. The relatively small sample size of 35 participants and the uncertainty of the quality of their responses could be contributing factors too.

Cronbach's alpha

Where the confirmatory factor analysis aims to assess construct validity, another necessary step is to assess construct reliability. The internal consistency of each construct can be measured using Cronbach's alpha [50]. This shows whether or not the ratings given by the participants on the survey questions are consistently correlated with each other, since this is not taken into account in the confirmatory factor analysis which aims to measure how well the statements compare on measuring the same construct. This value can range from 0 to 1, where 0 means there is absolutely no consistency between the ratings of the statements in a construct and 1 means there is perfect consistency and the statements are technically redundant. Generally, a value between 0.7 and 0.95 is deemed to be acceptable.

For Creativity, this value is 0.828, which is good. For Consistency, this value is 0.645, which is close to acceptable. For Fittingness, this value is 0.659, which is also close to acceptable. This means the ratings on these constructs are generally consistent, indicating that the statements are likely not ambiguous or superfluous.

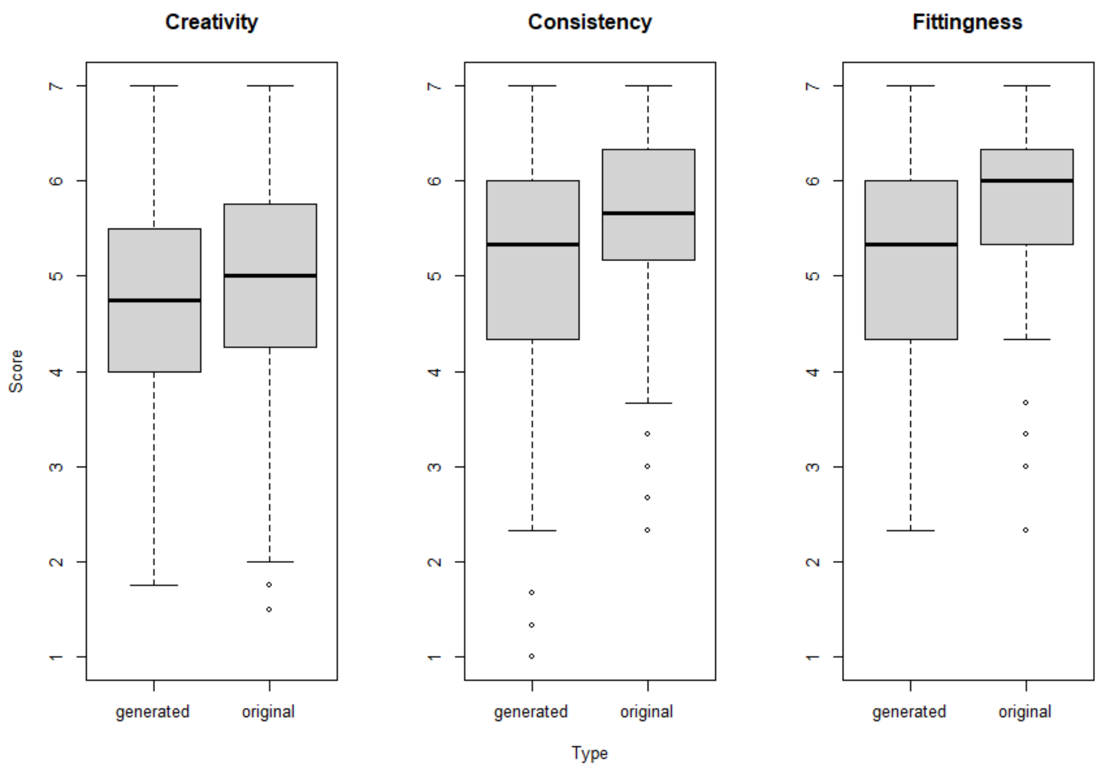

Figure 4: Boxplots of construct values of each construct, split by generated and original quests

5.2.3 Linear mixed-effects model

With the survey data, one linear mixed-effects regression models can be fitted for each construct, for a total of three of these models. By using this method, random effects (for example, the difference in how ratings are given out between participants) can be taken into account to estimate the difference in scores for each construct between the original and generated quest texts. The values of each construct's questions were averaged to create a construct value. These construct values were checked for normality of residuals and similarity of variances. The Q-Q plots and residuals can be found in Appendix B. If this would not be the case, the linear mixed-effects model would still be suitable to use for this data as these models are quite robust, but the certainty of the results would be somewhat reduced.

The distribution of ratings per statement is shown in Figure 5. Boxplots of these construct values can be found in Figure 4. Do note that the statistical comparison is made using estimated marginal means, while this boxplot figure simply shows the average construct values. Using the estimated marginal means takes into account the influence of the random effects as specified in the model equation below. This means the difference between the constructs reported from the linear mixed-effects model differs from the difference in scores as shown here. In this case, participants and quests were introduced as random effects, as some participants may be more critical than other, and some quests may be considered higher quality than others. Since these effects cannot be anticipated beforehand, they are taken into account as being random.

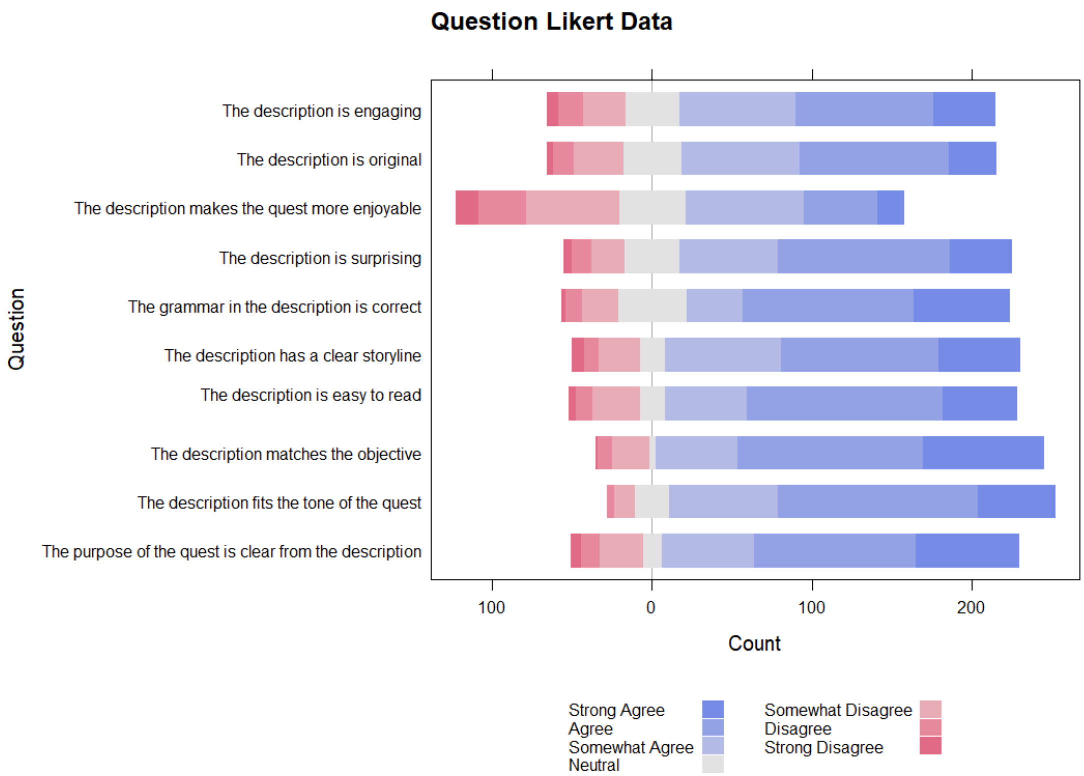

Figure 5: Distribution of Likert scale data per question

Each construct was used to fit a linear mixed-effects regression model with the following model equation:

$$Description\ Type \sim Construct\ Score + (1 \mid Participant\ ID) + (1 \mid Quest\ ID) \tag{4}$$

This means that the quest description type (which is either generated or original) is described by the construct score and that the differences among individual participants and quests are taken into account as random effects.

What follows are the results of the linear mixed effect models. In essence, for each construct, an average value is estimated for each type of description: generated and original. The generated descriptions are set as the intercept, which is essentially the zero point.

For Creativity, the model's total explanatory power is substantial (conditional $R2 = 0.35$) and the part related to the fixed effects alone (marginal R2) is 5.44e-03. The model's intercept, corresponding to generated descriptions, is at 4.76 (95% CI [4.34, 5.17], $t(275) = 22.65$, $p < .001$). Within this model, the effect of type original is statistically non-significant and positive (beta = 0.18, 95% CI [-0.07, 0.42], $t(275) = 1.44$, $p = 0.152$; Std. beta = 0.15, 95% CI [-0.05, 0.35]). This means the estimated average creativity score for the generated descriptions in this model is 4.76 on a scale from 1 to 7, and that this score is 0.18 lower than the original quest descriptions. However, this is not a significant difference at $p = 0.152$.

For Consistency, the model's total explanatory power is substantial (conditional $R2 = 0.39$) and the part related to the fixed effects alone (marginal R2) is 0.07. The model's intercept, corresponding to generated descriptions, is at 5.03 (95% CI [4.67, 5.39], $t(275) = 27.30$, $p < .001$). Within this model, the effect of type original is statistically significant and positive (beta = 0.57, 95% CI [0.35, 0.78], $t(275) = 5.21$, $p < .001$; Std. beta = 0.52, 95% CI [0.32, 0.71]). This means the estimated average creativity score for the generated descriptions in this model is 5.03 on a scale from 1 to 7, and that this score is 0.57 lower than the original quest descriptions. At $p < 0.001$, this is a significant difference.

For Fittingness, the model's total explanatory power is substantial (conditional $R2 = 0.27$) and the part related to the fixed effects alone (marginal R2) is 0.06. The model's intercept, corresponding to generated descriptions, is at 5.28 (95% CI [5.01, 5.56], $t(275) = 37.74$, $p < .001$). Within this model, the effect of type original is statistically significant and positive (beta = 0.48, 95% CI [0.27, 0.70], $t(275) = 4.42$, $p < .001$; Std. beta = 0.47, 95% CI [0.26, 0.68]). This means the estimated average creativity score for the generated descriptions in this model is 5.28 on a scale from 1 to 7, and that this score is 0.48 lower than the original quest descriptions. At $p < 0.001$, this is a significant difference.

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

5.2.4 Participant comments

There were some critical comments regarding the survey, both on the overall structure of the survey itself as it was deemed too long by a few participants, but also on the subject matter. One participant argued that it is not the descriptions that should be generated but that work on generating interesting objectives would make for more interesting quest design and would be a better use of the creative capabilities of language models. Concerns about the cookie-cutter structure of simple "fetch" or "kill" quests were raised, and the added value of generating the descriptions of these relatively simple quests were called into doubt. Some survey participants mentioned that none of the quests felt particularly good or bad, and that the objectives were generally quite unoriginal.

5.3 Discussion

From the results of both the confirmatory factor analysis and Cronbach's alpha, it can be observed that while this proposed structure of descriptive statements to describe these constructs is not perfect, it is quite close to being acceptable and should at least be suitable to draw some general conclusions from the survey data. Future research could take into account the results from this evaluation and use it to expand upon these concepts. The comparison of generated and original descriptions using linear mixed-effects regression models showed that the scores for Creativity, Consistency, and Fittingness were all lower by about 0.2, 0.6, and 0.5 points respectively on a scale from 1 to 7. While all scores are slightly lower for the generated descriptions, these are only significantly different for the measures of Consistency and Fittingness, and not for Creativity. Although these differences are relatively small and indicate some success with generating quest descriptions, several factors should be taken into account when interpreting these results.

Demographics

The survey participants mostly consist of young adults with about 90% of participants being between 18 and 34, and are mostly male at about 68% of participants. Comparing the experience of participants with the games used in this research, it can be observed that World of Warcraft is the most well-known among participants, with over half of participants having played it at some point, and all other participants having at least heard of it. The Elder Scrolls Online was second in this, with all but one participant having heard of the game and about a quarter of participants having played the game. The Lord of the Rings is close in results to The Elder Scrolls. Lastly, Neverwinter was considerably less known among participants, with almost half of participants having never heard of the game before and only a handful of participants having played it at some point.

The demographics present in this sample of participants do not show any specific outliers. Although only 36 participants actually finished the survey from a total amount of 78 people that started the survey, a completion rate of about 40% is not uncommon, especially considering the fact that there were some comments on the length of the survey. It is not exactly clear how many people were reached in total with the recruitment channels used, and as such the overall response rate can not be estimated for this survey. However, considering the multiple channels used, a relatively low response rate was likely. This may indicate a non-response bias, which means (part of) a demographic was not interested in filling in the survey. This could mean that potential participants are uninterested in the survey because it is not interesting to them or does not align with their views.

There are also more general concerns regarding the ethics of using people's creative work to train machine learning models. Video game development teams for these types of multiplayer RPGs can consist of hundreds of different employees, with dedicated teams for writing lore-accurate flavor text such as quest and item descriptions. Participants may consider the ethical implications of computational creativity, such as the potential of these methods to put people out of a job. This may in turn introduce a negative bias from participants towards these methods of quest writing. Another consideration is participant bias, which means the survey results could be biased towards more critical consumers of video games as these are more likely to participate in these types of surveys. Both of these facts may indicate the survey participants likely contains a sample of more opinionated players. This also means that any difference found between the original and generated quests could be more pronounced than it would actually be experienced by the average player during their gameplay experience.

Limitations

This evaluation method completely removes the quest from the original context. Although the ratings of the generated descriptions on the proposed constructs may be relatively similar, it is likely that many of these quests would not fit well when directly replaced in the game environment, as many of these quests are a part of overarching storylines. This is especially important for quests that are part of questlines, as earlier or later quests in such a questline may expect certain parts of the story to be present in these quest descriptions. Even if quests are not part of a questline, the quest completer may be involved in other quests and might have a completely different way of speaking to the player compared to the writing style that the model assigns to them. As such, the current evaluation results can only be interpreted in isolation.

Some participants' expectations of the type of quests in multiplayer role-playing are quite low, so it is hard to objectively say what a "good" description is. This expectation may be because of multiple reasons. One of these reasons is that because of the shared nature of these worlds, quests generally can not have a large impact on the game world, and can therefore not have stakes that are too high as the game world needs to be consistent for all players. Even if a game does have quests of relatively high quality or a quest line that constitutes a "main story", not all of the quests can be unique or contain high stakes.

The use of relatively simple quest structures and purely textual quests is also a limitation of this method, as more complex quest structures such as generating dialogue or entire quest lines, or an evaluation method with in-game implementation of these descriptions were out of the scope of this research. However, these factors may still be part of participants' expectations for quests in modern video games and may influence the participants' experiences with these descriptions significantly. Many modern single-player RPGs present their story through dialogue, and some of these games even allow the player to influence their gameplay experience by providing multiple options to progress through the dialogue and allowing players to experience multiple different story paths based on these choices. Comparatively, providing the backstory through a single paragraph of text may be considered old-fashioned and low-quality. The academic nature of the survey may also cause participants to be more critical than they would normally be, and experiencing these quests in a game environment might elicit completely different feedback from players than merely reading these descriptions in a survey.

6 Qualitative Evaluation

The generated outputs from the selected models were discussed during a semi-structured group interview for a qualitative and more detail-oriented evaluation. The information gained from the focus group interview can be used alongside the results from the survey to answer RQ1 and gain insights into shortcomings of quest design in multiplayer games to hopefully aid future research into this topic.

6.1 Methodology

The goal of the focus group interview was to obtain a more qualitative evaluation of the generated quest descriptions to be able to draw a stronger overall conclusion when combining these results with the statistical results from the survey. Another goal was to gain insights into the experience of quests in multiplayer RPGs in general with the hopes of identifying relevant aspects of quests that could be improved upon through the use of text generation. The focus group interview was set up as a semi-structured group interview with the opportunity for participants to have an open discussion among themselves regarding storytelling structures in multiplayer RPGs.

Recruitment and procedure

Participants were recruited through the survey. Survey participants were able to indicate whether or not they were interested in attending a follow-up focus group interview by leaving their email address at the end of the survey. These interested participants were then invited to this focus group interview.

At the start of the interview, participants filled in a consent form where they agreed to have the audio of the session recorded. This consent form can be found in Appendix A. After this, an open discussion was started, where the researcher asked a question and allowed the participants to answer this question and react to each other's answers. The researcher prepared a list of open-ended questions and allowed the discussion between participants to continue from there. If a silence would fall or participants felt that the question was adequately answered, either a follow-up question on the same topic or a question on a new topic would be posed. After these questions, the quests as seen in the survey were shown in a slideshow with the original and generated descriptions displayed side by side for comparison. Since every participant had seen either the original or the generated version of each quest, a discussion could be held on the participants' first impressions of these descriptions during the survey, and on their comparison now that the participants could see both versions.

The questions covered participants' experience with multiplayer RPGs, their opinions on the storytelling elements in these games, as well as questions on possible player motivations for playing through storytelling aspects of both single-player and multiplayer RPGs, and questions on the difference between the original and generated quest descriptions used in the survey. Discussion on the ethics of using language models for the purpose of creative text writing was explicitly left for the end of the session if there would still be time left. Although this is an important discussion on the topic of text generation, input on this topic was considered to be less useful at this time than the participants' input on the currently obtained results.

6.2 Results

Four participants were able to attend the focus group. These were all fellow or former students from the university. There was a variety of multiplayer RPGs that participants were either playing currently or had played in the past. These titles were World of Warcraft, TERA, Final Fantasy 14, Runescape, The Elder Scrolls Online, The Lord of the Rings Online, Sherwood Dungeon, Maplestory, and Aura Kingdom.

6.2.1 Open discussion

Because of the different methods of storytelling approaches used in these games and the fact that not every participant had played or was familiar with every different game mentioned, it was hard to compare the participants' experiences with the storytelling aspects in these games directly. Some of these games have a mandatory "main" storyline that dictates the player's progression through the game world, while others only have optional quests spread around the world that players can completely ignore if they choose to. Some of these games present their quests through pieces of text, while others present them through voice-acting, and yet others contain fully animated cutscenes. Despite these differences, several common experiences and discussion points still came up.

Storytelling experiences

When asked how the participants experienced the storytelling elements in these multiplayer games, several points were discussed. One was the lack of personal stakes and the lack of impact seen in the game world when completing these quests. Although some quests may radically improve the lives of the characters involved, this is generally not reflected anywhere in the game world. Often a location on the player's map is given as soon as the quest is accepted and as such, the quest's backstory becomes mostly irrelevant. Players do not need to read the quest text in order to know where to go, as it is already pointed out to them.

Another point was the lack of mystery: the assignment from each quest is clear from the single objective line that shows up in the quest log. Not many of these games implement a story that develops over time as the player progresses through it, so reading the quest text is not necessary. A positive example that was mentioned by a participant was a quest in The Lord of the Rings Online, in which the player is tasked to slay boars in a set amount of time. However, it turns out there are no boars in the area and as such, the player's expectations are subverted and the player is forced to return unsuccessfully after which the quest line progresses. Such a surprise can make a quest (and consequently, its backstory) more memorable, as the developers used the player's expectations of the quest to positively surprise them.

Lastly, the participants noted a lack of personal challenge or personalization provided by these quests. This reduced the interest in these stories even more. Every player receives the same quests, and is able to complete just about every quest without too much effort (aside from quests that are specifically group-focused as these are impossible to complete as a single player). There are generally no negative consequences for failing a quest aside from a small time loss for the player, and as such, the perceived stakes are also lowered for the player. While challenging content that also contains storytelling elements does exist in these games, it is generally not present in quests.

Improvements

When asked what aspects could be altered to improve the experience of quests in these multiplayer games, the story content itself was only a small part of the discussion. Aside from the previously mentioned points of quests having impact on the game world and quest having a stronger sense of mystery, participants also mentioned quest text personalization based on the player character's background, such as their race, class, background, or faction. Although some games already implement this in the form of tags (such as the <race> or <class> tags that are replaced with the appropriate information based on the player character in the World of Warcraft and The Lord of the Rings Online datasets), the rest of the text still remains the same and as such, this is not as interesting to players as completely different quest texts based on this character information.

Expanding on this beyond changing just text could be having different quest lines based on a player character's background. Games like World of Warcraft implement this somewhat: depending on the player race, players start in several different locations and are eventually all directed by converging quest lines into the few different routes of progression through the world. Another example already in use is unique quest lines based on the player character's class. However, examples like these could be further expanded upon using these text generation methods.

Quest text that changes based on previous quests done was also discussed, as this would increase the feeling of immersion into the game world. Multiple ways of coming into contact with the quest such as multiple quest-givers or game characters that point to the quest giver would be a welcome addition as well. Although a single non-player character wants something done and provides a quest, this could mean that other non-player characters know about this task as well and refer the player to this quest giver. An example mentioned by a participant was The Elder Scrolls Online. Quests in this game can often be started from multiple characters that are somewhat involved with the quest.

Another issue mentioned was the disconnect between solo and cooperative play. Being able to fully immerse into a story is generally opposed to being able to play together with other players while socializing. As such, players will have different expectations based on the context of their current gameplay session. A solution for this could be reducing the amount of backstory that is provided when a player is currently in a group with other players. Although making two versions of the same quest is usually not possible for developers, text generation could provide a solution here.

Manually setting up adapting quest texts to these kind of examples of adaptive quest structures would be time-consuming to realize for developers because of the large amount of different possibilities that each scenario introduces, but participants noted that text generation could be a solution to potentially reduce the development effort required.

Comparison to single-player games

When asked to compared quests in multiplayer RPGs to single-player RPGs, participants mentioned the difference in focus of each game's design. Namely, single-player games allow for a solitary, immersive experience where the player is an important character in the game world and the story is shaped by their choices and actions. Comparatively, although multiplayer games often refer to the player as being "special", it becomes clear from the gameplay that nothing the player does actually impacts the world all that much.

Development limitations

Some trade-offs that will remain challenges for developers were also recognized by the participants. These trade-offs will remain relevant for developers as they are rooted in a fundamental difference in design philosophies and/or player expectations. That does not mean there are no solutions for these problems, but that these trade-offs should be taken into account when considering the use of text generation. These considerations were challenge and accessibility: More challenging content will automatically be less accessible for unexperienced, low-skilled, or disabled players. This does not only encompass difficulty in gameplay, but also in the amount of details provided by a quest. Although a sense of mystery during quests may improve the experience for some players, the lack of clarity may reduce the experience for other players. This trade-off between clarity and mystery should be taken into account when considering text generation for quests.

The second trade-off concerns quest consistency: Although "uniqueness" of a quest is often named as a positive aspect, not every quest can be different or unique as this requires a baseline expectation to compare quests to. If there are no "normal" quests to compare to, these unique quests become the norm. As (the current implementation of) text generation models are generally quite similar and generic in output, they would likely be more useful for "standard" quests and not for more unique quests. This can be seen in the difference between the original and generated descriptions of Quest 2 from the World of Warcraft model.

Future expectations

When asked about the future expectations for quests in these games, participants mentioned an increasing focus on a "main" storyline for all players to experience, as well as an increasing focus on social play. Games are more likely to focus on horizontal progression. In essence, horizontal progression entails having multiple gameplay options, none of which are strictly "better". These different gameplay systems simply provide different experiences for the player to enjoy, and allowing players to play together more easily without being separated because of a difference in progression. This is in contrast to the previous focus on vertical progression, which is a gameplay system focused entirely on a single line of progression. This type of system is most commonly seen as players continually increasing their character's power through character levels and items to be able to enjoy increasingly challenging group content. However, participants mentioned that the impact and role of generative models in this transition was not yet clear for them.

6.2.2 Quest comparisons

When comparing the original and generated quest descriptions as seen in the survey, participant noticed a more chaotic, free-form writing structure in the generated descriptions. A quest giver's personality would often be more pronounced, such as the absent-minded professor in Quest 1 and the goblin accent in Quest 3 from World of Warcraft. The mystery in generated descriptions was often greater, with fewer facts about the quest mentioned or at least less directly alluded to. This was experienced both positively and negatively, as the quests would be more interesting, but also less clear. When fewer details were given in descriptions, these were sometimes described as giving the participants a sense of uncertainty. For these more mysterious descriptions, the player is expected to pay more attention to the context. However, since context is given in this format to estimate this impact, it was hard to estimate for the participants whether this would be positive or negative when experienced in the game.

Aside from the wrong factual information in some of the generated descriptions (such as cardinal directions or descriptions of locations), generated quests were considered to be very hit-or-miss in their quality and writing style. Participants mentioned that based on the context and placement of the quest, their expectations would be very different. Although a game's main story is expected to have consistently high quality and coherent writing, smaller, self-contained side quests such as these can afford to be less serious and more spontaneous in their story as the player's expectation is lowered. One example of this was mentioned for Quest 1 from World of Warcraft. If the quest giver would be standing near these creatures as if they were in pursuit, it would make more sense that they are distraught and consequently say things the way they are worded now in the generated description such as "It is not too late" and "Now, what were we talking about?" However, if the quest giver would be standing in a town (as is the case for this particular quest), this way of speaking makes much less sense. As the participants of both the survey and the focus group interview did not experience the quest as it was meant to be seen inside of the game, they could interpret the quest in very different ways and have differing opinions on its quality. This was a shortcoming of the format used for the entirety of this research.

In comparison, the original descriptions were described as feeling more like the writer had a bullet list of points that should be mentioned in the quest text, and then wrote the rest of the text around it to fit this structure. Some participants even mentioned that the writing style felt like "design by committee", as there did not seem to be an original writing style left.

In some cases, neither the original or the generated description was clearly preferred by the participants, as both were either very vague without any additional context, or they differed in such a way that they were not really comparable. An example of this was quest 3 from The Lord of the Rings Online, as in this quest the objective is already quite long, and either description does not really add much to the objective, the descriptions even contain different assignments. This might be because the objective is significantly different from other objectives in the training set and does not contain a very clear assignment, and as such, it is hard for the model to estimate how this description should look. This could be a limitation of the dataset and an indication of the amount of information required to expect high quality generated descriptions.

Participants mentioned that even if the descriptions were sometimes not very good, they would still pursue the quest as the expectations and goals associated with the quest are generally not focused on the story but more on the progression it gives the player character, although this would naturally differ per player.

6.3 Discussion

The focus group interview showed that the experience of quests in multiplayer RPGs in general is quite negative. This extends to the storytelling aspect as well. Quests are generally seen as obstacles that are necessary for quick progression of the player character. This could be because most of these multiplayer games are focused on vertical character progression: getting to a higher level, getting stronger items, getting more and stronger attacks, etc. Many players consider the "real" content to be the most difficult (and consequently, most rewarding) group content that is generally only found at the maximum character level at the "end" of the game. This is compounded by the lack of meaningful choices that players can make during these quests and the lack of significant impact on either the character (aside from progression) or the game world as a result of completing these quests. The story aspect of these quest will then be considered amusing at best and annoying at worst, but most likely irrelevant.

Different types of quests fulfill different purposes in gameplay, and currently one of the most popular uses for quests in this type of game is simply progressing the player's character (for instance through items or experience points). In this case, the story provided by the quest is not very important compared to the progress it brings the player. Another popular use is to provide a challenging experience that players can play through together, which also does not require a particularly strong storytelling element as the focus is on the cooperative gameplay. Although some points in this discussion are not directly applicable to text generation, they should still be taken into account if creative text generation is to be taken seriously as a consideration by game developers. Video game quests do not exist in a void, and changes made to the gameplay systems that incorporate them will impact the experience of these intricately designed game worlds. The design trends and conventions in these gameplay systems and their accompanying storytelling structures shape the expectations of players.

This was in contrast to how the experience of quests in single-player games was described by the participants, where the start of a quest is more often expected to be a setup for a larger story that develops over time and where much less of the quest is known at the start. Additionally, these quests often allow players to make choices that can then impact the development of the game's story. These games are not focused on getting to the end as fast as possible to play the "best" content, but instead allow the player to take their time with immersing themselves in the game world.

Some of the points mentioned are simply restrictions of the multiplayer RPG genre. Having multiple versions of areas based on a player's quest progression would be hard to justify towards game developers as it means double the effort of developing an area for something that would also split the player base into two: players that have completed that certain part of the story and players that have not may not be able to play together anymore if their game world looks significantly different. Because of this, quest stakes are forced to remain low for all quests. Additionally, fully personalized quest objectives is likely impossible in a world that thousands of different players all experience at once. The most likely avenue of personalized content in multiplayer quests would be to have differing texts that take the player's current situation into account but that would also result in similar objectives so that the overall experience for every player would be the same but the context in which it plays out could still be diverse. For instance, if a player is a human warrior, a human farmer that provides a quest could be more positive towards them than they would be towards an elf mage, and they might ask the elf mage to provide some sort of additional proof of their noble intentions. A good example of this that was mentioned by a participant can be found in the single-player RPG The Elder Scrolls 5: Skyrim. In this game, non-player characters will insult the player character if they choose to play as a Dark Elf, as this is the same race that aggressively occupies the region. Similar personalization is generally not yet found in multiplayer RPGs, but could still be part of survey participant expectations.

7 Conclusion & Future work

This chapter will discuss the findings and implications of this research, as well as answer the research questions, evaluate the research method by reflecting on possible shortcomings and limitations, and expand on possible avenues for future work.

7.1 Summary

The goal of this research was to generate backstories for quests in multiplayer online RPGs similar in quality to human-written quests. Previous research regarding this topic was discussed in Chapter 2, which showed that while some attempts have been made in automating the writing of these quest descriptions, it is still very much an active challenge. One promising method of doing this involves fine-tuning a large language model using a purpose-made dataset of quests from the game in question.

To do this, several video game quest datasets with pairs of objectives and descriptions were selected and gathered in Chapter 3. These datasets were from the multiplayer RPGs World of Warcraft, The Lord of the Rings Online, Neverwinter, and The Elder Scrolls Online. A selection of models was fine-tuned with one of the datasets and compared to choose the most suitable. The large language model used for the evaluation was decided to be GPT-J, as this was the largest and newest model that was still possible to fine-tune with the available resources as well as being easily available in the programming libraries used. This base model of GPT-J was then fine-tuned as described in Chapter 4. This led to four different models, namely one for each full game dataset, as well as four additional models for the largest datasets. For these four additional models, modified datasets were created where the meta-information was removed or a random selection of 1000 quests was made. This allowed for direct comparisons of the generated results from the full dataset to results from these partial datasets to estimate the influence of this meta-information and the size of the dataset respectively.

Generation results

From these generated descriptions, some preliminary observations were made regarding the quality of generation. One of the biggest limitations of this method is that the generated result is limited to the structure provided by the dataset, and it is not possible to steer the text generation in such a way that certain relevant topics such as characters or locations are always mentioned correctly. It cannot be assumed that a single generation result is suitable and as such, it still takes quite some effort to select the best result from a group of generated results. While the writing style and use of contextual information of most generated descriptions is quite good, the datasets and training structures for these models still need to be evaluated and adapted further to be able to consistently write both interesting and relevant backstories. Ideally, a user should be able to provide certain context about the quest that should be mentioned in the description and have the language model incorporate this into the description seamlessly. While this was attempted in the form of meta-information such as providing the quest giver or location and this does seem to aid the generation somewhat, it is still not enough to prevent mistakes in the generation. The motivation of a quest is not provided directly in the input but is only present in the example descriptions in the training set. This is compounded by the fact that these quests are generated in isolation, since questlines are also not directly referenced in descriptions. This results in the model needing to generate a fitting motivation without much information on the additional quests, leading to wrong assumptions when directly compared to the original description. This could also be due to the lack of detail present in the given meta-information. For example, the locations provided in the meta-information are entire zones while in reality, there are many subsections of these zones with their own unique properties. Other examples are the lack of scope or scale for each quest objective, and the lack of additional information on the involved non-player characters that may further inform on a suitable tone for the motivation such as the military rank or race of the quest giver.

The generated outputs of these models were then compared to each other, and the two best-performing models as decided by the researcher were chosen to be evaluated further with users. These were the models fine-tuned on the full datasets from World of Warcraft and The Lord of the Rings Online, as these showed the most promising and high-quality results.

User evaluations

The background research showed several different angles for defining computational creativity and tackling the evaluation of generated texts. As multiple papers advised a multifaceted approach to evaluation, a combination of evaluations was chosen in the form of a survey and a semi-structured focus group interview to provide both quantitative and qualitative data respectively. As there were no clear guidelines of evaluating this type of generated text and most previous research generally did not have a very expansive method of evaluation, the survey questions consisted of ratings on descriptive statements. These statements were ideated from keywords used to describe several factors of creative text writing, namely Creativity for assessing the creative value of the description, Consistency for assessing textual quality, and Fittingness for assessing the contextual appropriateness of the description compared to the given objective. Participants would randomly see either the original or generated version of each quest, and were asked to indicate their measure of agreement on these statements on a Likert scale of 7 points. These measurements on original and generated descriptions were then compared

The statistical results from the survey indicated no significant difference in Creativity scores but a significant difference of about half a point on a 7-point Likert scale in favor of the original descriptions for Consistency and Fittingness. Although these results are quite good, it should be taken into account that these generated quest descriptions are pulled out of context and would not fit naturally into the game if directly replaced. More strict requirements would be necessary to be able to use this kind of generation for quests that need to fit into a larger narrative such as a quest line consisting of multiple quests as is quite common in this type of game. While meta-information for quest lines was implemented into the datasets, this did not seem to consistently aid the generation results. This could be because the previous or following quests in a quest line are not provided in the context window when generating quest descriptions, and the fact that the complete information of each quest is not completely saved in the model weights either. While this could be possible in future implementations by providing previously generated descriptions as additional meta-information input for generating subsequent descriptions, it was considered out of scope for this research as it would require a more complex dataset structure.

The academic nature of this survey is not representative of the actual experience that players would have compared to playing through these quests in an actual video game environment. This may also have led the survey participants to be more critical end-users than average players. Multiplayer RPGs generally have their quests located and structured in such a way that it is very easy to accept multiple relevant quests for a certain region and play through them by yourself or meet other people along the way to play along with. The impact of a single backstory from a quest will therefore be less relevant than the broader gameplay experience provided by the quest in practice, and even implementing these quests into a game environment itself would not solve the difference between the academic testing environment and all of the possible different gameplay experiences that would naturally occur between strangers in a large and persistent virtual world. This method of storytelling through text could also be considered outdated by modern standards, as multi-modal presentation methods of stories becomes increasingly expected from developers. It is hard to estimate exactly how this influenced the results, but these factors should be taken into account when attempting to draw definitive conclusions. Taking all of these points into account, the estimated difference in reception of both the original and the generated quest descriptions is not very large when compared in isolation, especially on the measure of Creativity.

The focus group interview showed that there is a noticeable difference in writing styles between the types of descriptions, and that this is not necessarily experienced negatively as was presumed beforehand. Aside from noticing the factual errors when comparing the descriptions, participants mentioned that personality traits of the quest givers seemed to come forward more positively in the generated descriptions, as these were often more exaggerated in the writing style. Generated descriptions were considered to be more unpredictable and mysterious, as less information was provided to the player. Consequently, these descriptions were seen as more interesting to read than the original descriptions. The original descriptions were described as bullet point lists where some background information was added after the required facts of the quest were clear enough. This method of evaluating also provided insights into some of the limitations of quest design for multiplayer RPGs, as well as some potential implementations for feasible improvements.

7.2 Research questions

Research question 1 (RQ1) was presented as follows:

> To what extent can large language models be fine-tuned to generate backstories for multiplayer role-playing video game quests similar in quality to human-written quests?

Given the considerations from both the survey and the focus group interview, the answer to this question is therefore that it is not yet possible to fine-tune a language model to generate video game quest descriptions similar in quality to human written quests. While the generated descriptions are rated positively on average, these results are based on isolated instances of quest descriptions. The current amount of inconsistencies and errors in generated descriptions means the original descriptions could not be directly replaced with the current results while maintaining the same quality.

Many improvements can still be made in terms of making the descriptions fit more closely to the given objective and reducing the amount of hallucinations and errors. Some improvements were already made compared to previous research on this topic by incorporating newer language models, and using several differently structured datasets with meta-information embedded. Comparing the current results with the results from previous research does show some of the progress towards this goal, but much still remains to be done to be able to easily incorporate these techniques as a developer and trust the generated results.

Research question 2 (RQ2) was presented as follows:

> To what extent does the embedding of additional quest information influence the quality of quest story generation?

The generated descriptions from models trained on datasets with and without the additional meta-information were compared in Chapter 4.3. Since the objectives currently do not contain all of the required contextual information for a quest, the additional meta-information is observed to inform the model on the correct use of named entities, as well as on the sense of urgency and seriousness for each quest. Generated descriptions without meta-information show a lack of knowledge and a lack of concern for the tone of the story context, which translates to less interesting and less appropriate descriptions. The meta-information also seems to correctly inform the model on the appropriate writing style and tone for the different quest givers.

Although it was observed that context is important for the quality of results, the best method of correctly using this context is still under debate. This could be entrusted to the model such as in this research, or it could be done manually. This second option means replacing named entities by generalized tags in the training examples, teaching the model to generate these tags instead of directly using names, and replacing these tags with context-appropriate names after generation.

However, this would likely also reduce the chances of the model picking up on relational information between relevant entities and words that indicate the tone of quests. This method would also require the replacement of this contextual information in the dataset beforehand, and the replacement of these tags with this information after generation. In essence, by generalizing this training data through the use of tags, the uniqueness of the original descriptive wording will likely be lost. Compared to this option, the currently proposed method of supplying the relevant context in the input seems to preserve the original word use better and as such, leads to more genuine outputs.

Research question 3 (RQ3) was presented as follows:

> What is the influence of the amount of quests used in fine-tuning on quest story generation quality?

Although the models trained on a subset of 1000 random quests performed quite well and showed interesting results based on a relatively small amount of quests, some distinct differences between these results and the results from the full datasets could be observed. Since three descriptions were generated per quest and one of these was picked manually as the final result, an increase in variety in terms of motivation and quality was noticeable during this selection process. The results from the model trained on the full dataset were observed to differ only slightly in wording and background per quest, while this variety was much higher for the smaller dataset. In these results, all three descriptions often contained very different backstories and motivations for the same quest.

This increased variety has upsides and downsides. While more different types of backstories could be picked, this need for cherry-picking the desired result also increases the manual amount of work involved for the developer. The variety in quality also decreases the amount of confidence in any single output from the model. However, if used indirectly as an ideation tool for developers instead of intending to directly replace the original descriptions with generated descriptions, this increased variety may be more useful for developers.

7.3 Limitations and shortcomings

As discussed, there are some limitations to the current methods used. The biggest limitation of the currently used method is the reliance on the structure and description styles of the available datasets. Although multiple different datasets were attempted, two turned out to be relatively low in quality. However, even the better datasets do not contain quest motivations or many other restrictions for the content of the quest description. More work is required to set up datasets with these properties, but this additional required effort further reduces the viability of such a system for use in actual game development, which was a goal of this research.

This research was limited by the amount of available processing power, as well as the limited documentation on training and generation parameters in the PyTorch, HuggingFace Transformers, and DeepSpeed libraries. This required a lot of research time to be dedicated towards this experimentation, and overall made it more difficult to estimate model performance during training and generation.

Some oversights were made by the researcher during this entire process as well. These include the potential bias that was introduced by selecting the "best" result from three generated descriptions, as well as any mistakes that were missed during cleaning of the datasets that may have influenced generation results. The selection of different models tested is small at only three models, and evaluation of these models was done subjectively. Although the use of LLaMa-7B was also attempted, this was not possible at the time. Other decoder-only models such as BLOOM-3B, OPT-2.7B, or OPT-6.7B, or even encoder-decoder models such as T5-3B may have been worthwhile to investigate as well. Although attempting to use increasingly larger LLMs for this purpose might make sense at first, Ouyang et al. [7] showed the performance improvement of a smaller fine-tuned model of "only" 1.3B parameters over a standard version of GPT-3 at 175B parameters. As such, it is likely better to figure out the best method and dataset structure for fine-tuning than attempt to use increasingly larger (and computationally expensive) models.

The current evaluation methods are highly academic in nature, and consequently, the results can only be interpreted in isolation. The currently proposed evaluation measures require additional adjustments to be worth using for drawing well-founded conclusions from the survey data, and the current results only use a relatively small sample of participants. The representativeness of this sample is also debatable. Instead, this evaluation methodology using the descriptive statements in a linear mixed model should be considered as a proof of concept.

7.4 Future work

Considerations for future work include the collection and use of even more diverse, detailed, and higher quality datasets. A major oversight of the current implementation is that there is no way to set strict requirements for the contents of the description. This meant the model had to contextually link the terminology used in the training examples to the rest of the input that was provided such as the location or quest giver name. Further effort could be put into deconstructing the original descriptions using a method like Named Entity Recognition and adding this information in the training dataset as textual requirements for the generated description output. Other options are the addition of the required underlying motivation or tone for the quest. As discussed in Chapter 3.5, additional columns for motivation, game expansions, more detailed information on the quest giver, and further requirements for the description could make for more specific models and may improve generation results as a consequence. Some of the missing factual information could still be gathered from the sources currently used during gathering and would only require more complex scraping methods, but other information such as motivations or restrictions per quest would have to be set up manually. Using a random subset of 1000 quests showed an interesting balance between output variety, and dataset complexity and size, which might be interesting for further research.

Another consideration is the use of quantized models for lower memory consumption during training and inference. Most current models are trained and inferenced using 32-bit floating point numbers, but by using a lower-bit optimizer for fine-tuning and loading the model in a lower-bit mode for inference, memory consumption can be reduced. While this implementation uses half-precision 16-bit floating points, this could be further reduced to 8-bit or even 4-bit integers to vastly reduce the memory consumption during fine-tuning and inference. These data types are less precise, but for these types of models this reduction in accuracy is often quite small [51].

An in-depth comparison of several popular LLMs was considered at the start of this research but quickly proved to be outside of the scope of this research, as this would require a much more in-depth comparison between the various different LLMs that could be considered. Finding a good balance between model requirements and performance could prove beneficial for future research in this topic. Another part of this is to focus more extensively on the process of fine-tuning and performing a hyperparameter sweep to accurately assess the best fine-tuning and generation parameters for optimum performance of this text generation task. This was currently estimated manually by the researcher because of the time constraints of this research.

Similarly, research into the changes to the attention values in these models could be performed using tools like BertViz [52] to give insight into the effects of including contextual different pieces of meta-information. For instance, the current results do not show a clear use of the *Previous/required quest* or *Next/Unlocked quest* properties, but this is purely observational and further research would be needed to confirm or deny these observations.

The current XML-like bracket structure of the fine-tuning prompts as shown in Chapter 3 may also not be suitable. In this implementation, the models are also being trained to recreate this structure and generate the meta-information instead of only being trained to identify structures in this input data and only writing the descriptions given this context. For example, in its current form the model may become very good at predicting everything about the quest except the descriptions and still score quite high when calculating the training loss. Since the data cannot be assumed to be evenly distributed and unbiased, this may prove to be a problem during training. Different methods of setting up the text data for training and generation such as Värtinen's descriptive data structure [9] or a setup with a question-answer data structure may offer solutions for this issue and may prevent potentially fine-tuning the model in a wrong way. Other data structures such as a relational database such as the one used by Ashby et al. [29] could also be promising to pursue further, although the large scope of these types of multiplayer games may prove to be too complex to put into such a database.

There is also the distinction to be made between pre-generated quests and live-generated quests, where pre-generated quests are those that are generated beforehand and then implemented into a game such as in this research, and live-generated quests that are generated on the fly in-game. While the structure of pre-generated quests is more rigid and controllable, live-generated quest descriptions could theoretically be adapted more expansively to the player's current context. Proponents of the use of artificial intelligence for quest generation in games will generally be more enthusiastic about the latter because of its expanded possibilities for different experiences in gameplay. This implementation would be significantly more complex and require an increased requirement for available processing power and the creation of suitable datasets. However, it could be an interesting research avenue.

7.5 Contributions

This thesis evaluated previous research in the application of computational creativity in the form of fine-tuning an LLM for generating video game quest backstories. It shows the entire process of gathering and cleaning datasets and fine-tuning generative models with these datasets, as well as setting up an evaluation framework with specific appropriate measures for evaluating these results through inferential statistics. Solutions to the existing limitations of the application were proposed, executed, and evaluated, and the currently obtained results show improvements in textual quality and the use of contextual information compared to previous research. Additionally, many new limiting factors were identified, which indicate potential improvements for consideration in future research. By providing as much information and insights as possible on the previous and existing limitations as well as some potential future avenues for research, this work can hopefully serve as a stepping stone for future research and as an insight into some of the current challenges of computational creativity.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[2] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020. arXiv: abs/2001.08361.

[3] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," 2017. arXiv: abs/1704.01444.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, and P. Dhariwal, "Language models are few-shot learners," 2020. arXiv: abs/2005.14165.

[5] M. U. Hadi, Q. Al-Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects," 9 2023. techRxiv: articles/preprint/23589741.

[6] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 9, no. 1, pp. 1–22, 2013.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, and P. Mishkin, "Training language models to follow instructions with human feedback," 2022. arXiv: abs/2203.02155.

[8] J. van Stegeren and J. Myśliwiec, "Fine-tuning gpt-2 on annotated rpg quests for npc dialogue generation," in *Proceedings of the 16th International Conference on the Foundations of Digital Games*, FDG '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[9] S. Värtinen, P. Hämäläinen, and C. Guckelsberger, "Generating role-playing game quests with gpt language models," *IEEE Transactions on Games*, pp. 1–12, 2022.

[10] K. K. Yu, N. R. Sturtevant, and M. Guzdial, "What is a Quest?," in *Joint Proceedings of the AIIDE 2020 Workshops (AIIDE-WS-2020)*, 2020.

[11] G. Smith, R. Anderson, B. Kopleck, Z. Lindblad, L. Scott, A. Wardell, J. Whitehead, and M. Mateas, "Situating quests: Design patterns for quest and level design in role-playing games," in *Interactive Storytelling: Fourth International Conference on Interactive Digital Storytelling, ICIDS 2011, Vancouver, Canada, November 28 – 1 December, 2011. Proceedings*, (Berlin, Heidelberg), p. 326–329, Springer-Verlag, 2011.

[12] K. K. Yu, M. Guzdial, and N. R. Sturtevant, "The Definition-Context-Purpose Paradigm and Other Insights from Industry Professionals about the Definition of a Quest," in *17th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2021*, pp. 107–114, 2021.

[13] A. Machado, P. Santos, and J. Dias, "Towards a Procedurally Generated Experience: A Structural Analysis of Quests," in *Proceedings of Videojogos 2016, Covilhã, Portugal*, 2016.

[14] P. Landwehr, J. Diesner, and K. Carley, "The words of warcraft: Relational text analysis of quests in an mmorpg," *Breaking New Ground: Innovation in Games, Play, Practice and Theory - Proceedings of DiGRA 2009*, 05 2011.

[15] E. S. de Lima, B. Feijó, and A. L. Furtado, "Procedural generation of branching quests for games," *Entertainment Computing*, vol. 43, p. 100491, aug 2022.

[16] C. Garbacea and Q. Mei, "Why is constrained neural language generation particularly challenging?," jun 2022. arXiv: abs/2206.05395.

[17] C. Gârbacea and Q. Mei, "Neural Language Generation: Formulation, Methods, and Evaluation," 2020. arXiv: abs/2007.15780v1.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. OpenAI blog.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, (San Diega, CA, USA), 2015.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. arXiv: abs/1711.05101.

[22] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," 2019. arXiv: abs/1909.05858.

[23] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 889–898, Association for Computational Linguistics, July 2018.

[24] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020.

[25] C. Meister, T. Pimentel, G. Wiher, and R. Cotterell, "Locally Typical Sampling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 102–121, 01 2023.

[26] J. Maynez, S. Narayan, B. Bohnet, and R. T. Mcdonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[27] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," 2020. arXiv: abs/1905.12616.

[28] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2019. arXiv: abs/1909.08593.

[29] T. Ashby, B. K. Webb, G. Knapp, J. Searle, and N. Fulda, "Personalized quest and dialogue generation in role-playing games: A knowledge graph- and language model-based approach," in *Proceedings of the 2023 CHI Conference on Human Factors in*

Computing Systems, CHI '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[30] J. Doran and I. Parberry, "A prototype quest generator based on a structural analysis of quests from four mmorpgs," in *Proceedings of the 2nd International Workshop on Procedural Content Generation in Games*, PCGames '11, (New York, NY, USA), Association for Computing Machinery, 2011.

[31] A. Machado, P. Santos, and J. Dias, "A procedural quest generator for Conan Exiles," master thesis, Instituto Superior Técnico, University of Lisbon, 2021.

[32] V. Breault, S. Ouellet, and J. Davies, "Let CONAN tell you a story: Procedural quest generation," *Entertainment Computing*, vol. 38, pp. 1–10, 2021.

[33] E. S. de Lima, B. Feijo, and A. L. Furtado, "Procedural Generation of Quests for Games Using Genetic Algorithms and Automated Planning," in *18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, vol. 2019-Octob, pp. 144–153, IEEE, oct 2019.

[34] P. Ammanabrolu, W. Broniec, A. Mueller, J. Paul, and M. O. Riedl, "Toward Automated Quest Generation in Text-Adventure Games," in *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, pp. 1–12, sep 2019.

[35] A. Alvarez and J. Font, "Tropetwist: Trope-based narrative structure generation," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22, (New York, NY, USA), Association for Computing Machinery, 2022.

[36] J. Van Stegeren and M. Theune, "Fantastic Strings and Where to Find Them: The Quest for High-Quality Video Game Text Corpora," *Joint Proceedings of the AIIDE 2020 Workshops (AIIDE-WS-2020)*, vol. 2862, 2020.

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[38] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[39] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Computer Speech Language*, vol. 67, p. 101151, 2021.

[40] M. Hämäläinen and K. Alnajjar, "Human Evaluation of Creative NLG Systems: An Interdisciplinary Survey on Recent Papers," in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 84–95, 2021.

[41] A. Jordanous, "A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative," *Cognitive Computation*, vol. 4, no. 3, pp. 246–279, 2012.

[42] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, and A. Moi, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, and G. Chanan, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[44] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," Mar. 2021. HuggingFace Transformers - GPT-Neo-2.7B.

[45] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," May 2021. HuggingFace Transformers - GPT-J-6B.

[46] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, and D. Hesslow, "Bloom: A 176b-parameter open-access multilingual language model," 2023.

[47] S. Wang and P. Kanwar, "BFloat16: The secret to high performance on Cloud TPUs," 2019. Google Blog.

[48] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," 2023. arXiv: abs/2308.08747.

[49] D. Hooper, J. Coughlan, and M. Mullen, "Structural equation modeling: Guidelines for determining model fit," *The Electronic Journal of Business Research Methods*, vol. 6, 11 2007.

[50] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International Journal of Medical Education*, vol. 2, pp. 53–55, 06 2011.

[51] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," 2017. arXiv: abs/1712.05877.

[52] J. Vig, "Visualizing attention in transformer-based language representation models," 2019. arXiv: abs/1904.02679.

A Consent forms

Consent Form for Survey on Quality of Quest Texts

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated 16-05-2023, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves rating several video game quests on measures such as creativity, coherence and fittingness in context in an online survey environment filled in by me.	○	○
Use of the information in the study		
I understand that information I provide will be used for a master thesis that will be made publicly available on the University of Twente Thesis database.	○	○
I understand that if I decide to share my contact information for an optional follow-up interview, this personal information will not be shared beyond the study team.	○	○

Click on the arrow button below to continue to the next page.

(The online questionnaire will have a requirement for participants to tick all boxes "Yes" before they can click the button to continue to the next question)

UNIVERSITY OF TWENTE.

Consent Form for Interview on Quality of Quest Texts

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated 16-05-2023, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves rating and discussing the text of several video game quests on several different measures such as creativity, consistency and fittingness. The audio of this interview will be recorded for the sole purpose of this study.	○	○
Use of the information in the study		
I understand that information I provide will be used for a master thesis that will be made publicly available on the University of Twente Thesis database.	○	○
I understand that personal information collected about me that can identify me, such as my name, gender or prior experience with video games will not be shared beyond the study team.	○	○
I agree to be audio/video recorded.	○	○

Signatures

______________________ ______________________ ________

Name of participant

Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

________________________ ___________________ ________

Researcher name Signature Date

UNIVERSITY OF TWENTE.

B Construct Q-Q plots and residuals plots

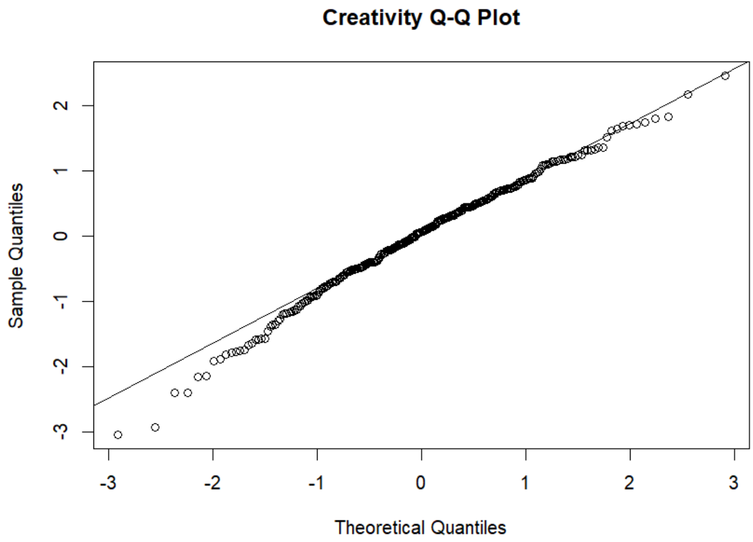

Figure 6: Q-Q plot of Creativity construct

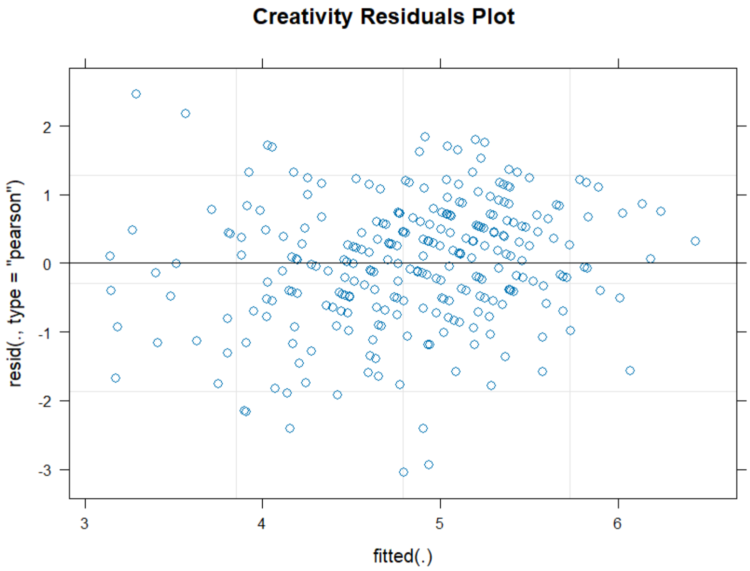

Figure 7: Residual plot of Creativity construct

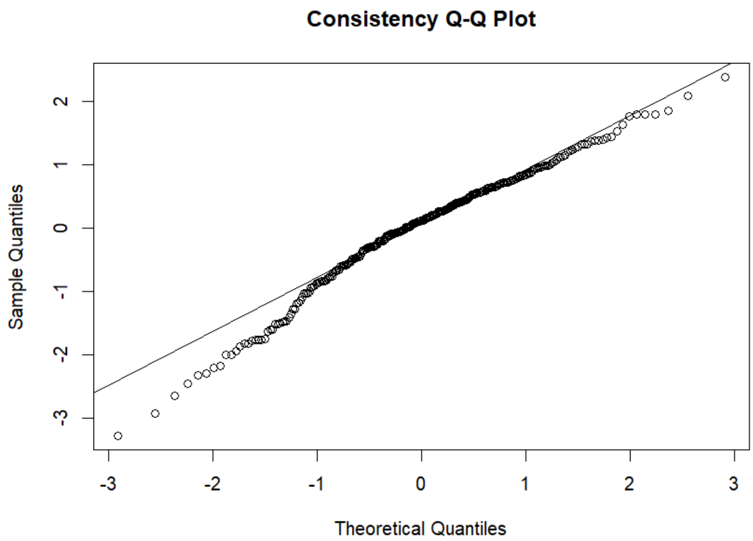

Figure 8: Q-Q plot of Consistency construct

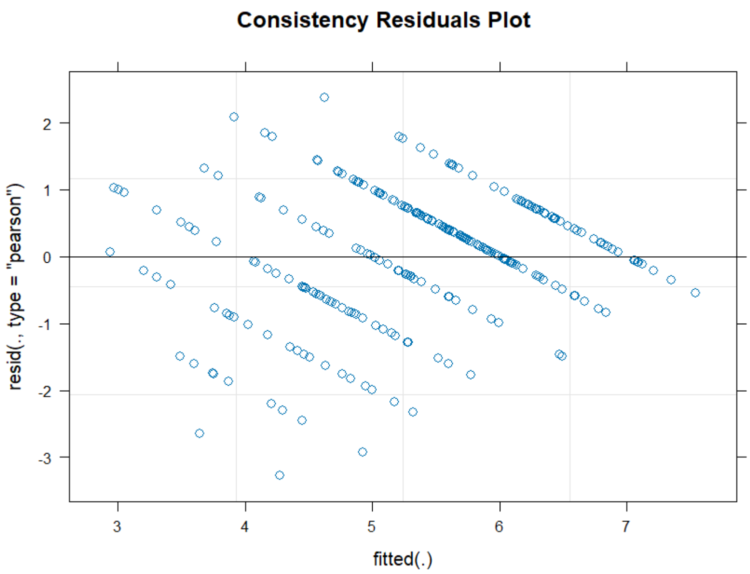

Figure 9: Residual plot of Consistency construct

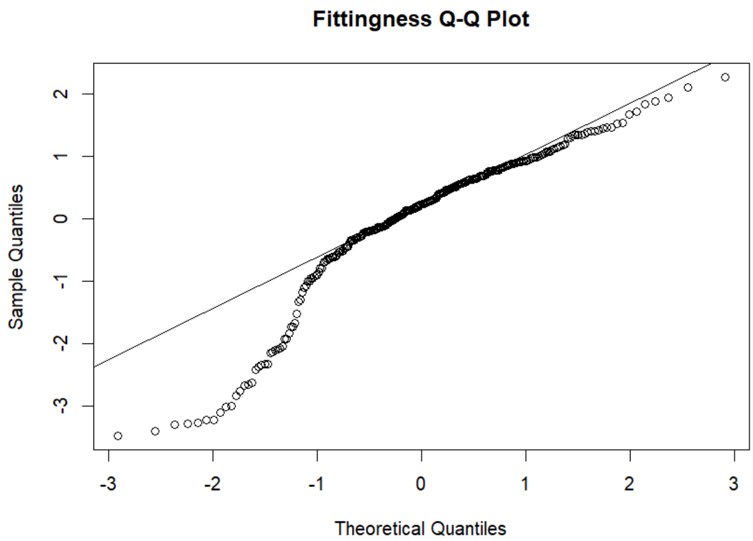

Figure 10: Q-Q plot of Fittingness construct

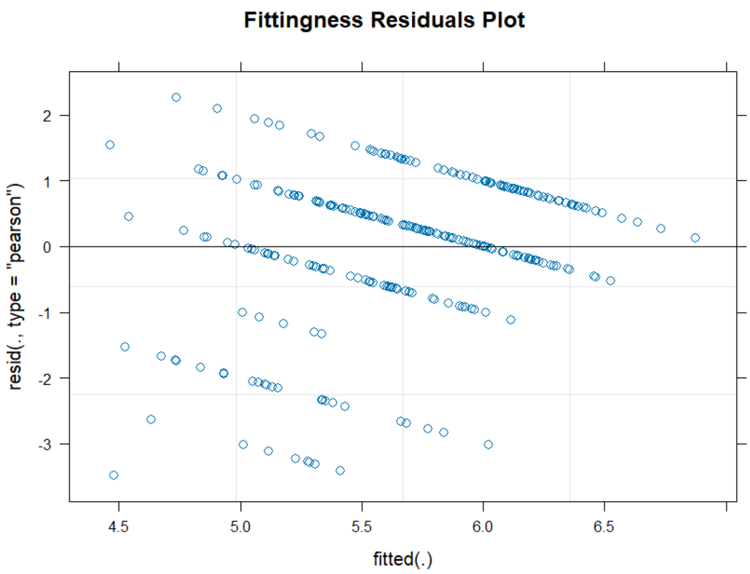

Figure 11: Residual plot of Fittingness construct