Fine-Tuning GPT-2 on Gameplay Data for Developing Context-Specific Storyline Generation

Andre Zhan Hong - XDST6, Jerry Pan - ZRMH8, Jay Wu - ZWQY2,
Kayla Yang Zhang - ZBGH0

University College London, School of Management
MSIN0221- Natural Language Processing

Word Count: 3630

Table of Contents

1. Introduction

Natural Language Generation (NLG) is widely acknowledged as one of the most complex computational tasks within the realm of Natural Language Processing (NLP) problems (Sidu Lu et al., 2018). Notwithstanding the inherent challenges, NLG techniques have demonstrated great potential in a variety of industries, particularly in the context of role-playing games, which often prioritise narrative-driven gaming experiences. This genre often includes the lore of the game (game world backstory and history of the characters), as well as side-quests, which are optional tasks designed to give players more flexibility. Given their typical structural consistency, character biographies are prime candidates for procedural generation. This study aims to conduct an exploratory analysis on the application of NLG techniques in the procedural generation of game character biographies as well as potential game storyline, using the popular online multiplayer role-playing game, League of Legends, as a case study.

1.1. Literature Review

In recent years, deep learning and pre-training models such as ELMo, OpenAI GPT, GPT-2, and BERT have exhibited outstanding performance in a wide range of natural language tasks. Compared to the traditionally used recurrent neural networks (RNNs) in NLP tasks, which can be computationally expensive given the sequential nature of RNNs which unables parallelised training; transformer models present a solution to this issue by avoiding the use of RNNs and convolutional neural networks (CNNs). Instead, they rely solely on attention mechanisms, making text generation models much less computationally expensive (Ressmeyer et al., 2018?). The improved efficiency of these transformer-based language models facilitates pre-training on extensive unlabelled text datasets. Consequently, fine-tuning pre-trained models has emerged as the best practice for achieving state-of-the-art results. GPT-2, in particular, has been recognised for its superior text generation quality when compared to BERT (Lewis et al., 2019).

In this project, we employ reinforcement learning-based text generation methods using OpenAI's GPT-2 source code and fine-tuned the medium-sized model for generating game story-play. While GPT-2 demonstrates efficacy in text generation, its performance can be enhanced by fine-tuning on a specific task. This involves using pre-trained weights and further training the model on a smaller task-specific dataset. Transfer learning has become an increasingly popular technique in NLP, and for GPT-2, fine-tuning allows the model to learn patterns and characteristics unique to the target task, resulting in improved performance. We were impressed by the coherency and complexity of our generated text; however, the quality varied. Surprisingly, only a few training steps were required for GPT-2 to generate text resembling storylines. Given the recent release of the most powerful model, GPT-4, we believe it essential to explore the potential applications of NLP text-generating models, their impact on the creative gaming industry, and how best to optimise their usage for gaming text generation.

Nonetheless, our study has limitations, including the surface-level coherency of the generated text, which requires further qualitative and quantitative analysis to assess the meaningfulness of the generated claims at a larger scale. We hypothesise that training GPT-2 on carefully selected data will yield more meaningful results; for instance training the model on data specific to our chosen game 'League of Legends' could reduce the generation of irrelevant text. Further research is needed to explore this hypothesis, given our study's limitations.

1.2. Objectives

The objectives of our project are threefold:

1. Our primary objective is to develop an auto-generated story model by using champion and region descriptions from the video game League of Legends. To achieve this, we will employ the GPT-2 pre-trained model, refining and fine-tuning the model to the text data.
2. Our second objective is to investigate the capabilities of transformer architectures, specifically GPT-2, in text generation and examine their adaptability to our specific task. We will monitor the generated text at each step of the fine-tuning process and track the loss values until convergence is reached.
3. Lastly, we aim to evaluate the overall quality of the generated text using various evaluation metrics. As we move forward with this phase, we will employ BERT as our metric to assess our progress. By comparing the generated text with the transcript of Arcane, a League of Legends-based animated series, we can gauge the similarity between the two and subsequently evaluate model performance.

Through this project, we aspire to evaluate the potential benefits our findings may offer to researchers, game developers, publishers, and players.

1.3. Business Application and Project Relevance

The proposed model holds particular relevance for businesses requiring auto-generated story models, such as the gaming industry, where engaging narratives are important for player immersion. A notable example of neural text generation in the video game industry is AI Dungeon, a hybrid between a classic text adventure game and a virtual tabletop Dungeon Master (REF). AI Dungeon was created by fine-tuning a pre-trained GPT-2 model on a generic dataset, similar to our approach. On the contrary, our model aims to automate the generation of new content, in particular character descriptions and storylines, serving as an authoring aid for game writers. This will save businesses time and resources, allowing them to allocate efforts to other aspects of game development.

Other industries, such as Advertising and Marketing, Film and Television, and Education, may also benefit from our findings. Advertisers and marketers frequently employ storytelling to establish emotional connections with their target audience. In the film industry, screenwriters and filmmakers could also leverage the model to rapidly generate plot ideas, character descriptions, as well as film dialogue. In education, teachers and educators may also benefit from an auto-generated story model, so as to use storytelling to engage their students and convey complex concepts.

Creating captivating narratives is a crucial aspect of game development, as it fosters user retention and attraction. Nevertheless, crafting these narratives and storylines is often time consuming and costly. Automating this process can improve time efficiency, and auto-generated story models can enable businesses in the gaming industry to generate engaging content quickly and effectively. This improvement not only streamlines the content creation process but also elevates the overall user experience, resulting in greater user engagement and retention.

2. Data Pipeline

To fine-tune GPT-2's pre-trained model in alignment with our project objectives, we established a data pipeline consisting of three stages: (1) raw data collection, (2) data pre-processing and cleaning, and (3) training set labelling.

2.1. Data collection

The text corpus was extracted from the description of all champions from League of Legends. Descriptions are available at: League of Legends: Champion Descriptions.

2.2. Data Pre-Processing

After collecting the necessary corpora, the texts are pre-processed and transformed to fit into a tabular format dataset. To do so, we defined a function that splits the descriptions by punctuations and removes trailing spaces, as well as special characters. The cleaned sentences are then transformed into a data-frame for modelling. Additional pre-processing techniques such as stop-word removal and lemmatization were not conducted, as our model already takes into account.

2.3. Training/Dev/Test Sets Split

The dataset was split into training, validation and test sets. The splitting process was distinct from the traditional method, and was as follows:

- Test set was selected as the last 20% of the completed dataset.
- From the remaining 80%, we used the train_test_split function to split 75% into training and 25% into validation.

3. Methods: Experimental Setup and Model Development

In this section we discuss our computational environment and set up; our code base, as well as the models we used to fine tune GPT-2. In terms of computation setup, we leveraged Google Colab's jupyter notebook environment. The cloud environment serves as a method to overcome memory constraints, as well as the compatibility to the transformers API.

After fine-tuning GPT-2, to build our text generator model, we proceeded into the following two stages:

1. In the first stage, we utilised the champions' stories featured in Arcane as our primary data to train our model. Arcane is a popular animated television series produced by Riot Games, the creators of League of Legends. The series is set in the world of Runeterra, which is the same universe as League of Legends.
2. In the second stage, we expanded our dataset, which will take every champion in League of Legends in our dataset for training and modelling. The result and output will be evaluated by human evaluation metrics. Our expectation is to create accurate auto-generated text for the storyline.

Model 1: GPT-2 as Baseline

GPT-2 is one of the powerful pre-trained language models developed by OpenAI and is widely used as a baseline model for various NLP tasks. With 1.5 billion parameters, GPT-2 was trained on a large corpus of text data, consisting of roughly 45 terabytes, including books, academic papers, Wikipedia and the internet, which enables such a pre-trained model to generate high-quality texts.

Using GPT-2 as the baseline model provides us with a benchmark for evaluating the performance of a fine-tuned model in the context of our specific task. The fine-tuning process involves employing transfer learning, where the reference model is initially trained on character and region descriptions from 'League of Legends', in order to facilitate

	champion	sentences
26	garen_27	When the new Sword-Captain of the Dauntless Vanguard fell in battle, Garen found himself put forward for command by his fellow warriors, and the nomination was unopposed.
27	garen_28	To this day, he stands resolute in the defence of his homeland, against all foes.
28	garen_29	Far more than Demacia's most formidable soldier, he is the very embodiment of all the greatest and most noble ideals upon which it was founded.
29	fiora_1	As the youngest daughter of the noble Laurent family, Fiora seemed destined for a life as a political pawn, to be married off in Demacia's grand game of alliances.
30	fiora_2	This did not sit well, and from an early age she deliberately defied every expectation placed upon her.
31	fiora_3	Her mother had the finest craftsmen of Demacia fashion, the most lifelike dolls for her to play with—but Fiora gave them to her maids, and took up her eldest brother's rapier, forcing him to give her lessons in secret.

Figure 1: Snapshot of the dataset.

adaptation and generation of specific text. This approach involves updating the pre-trained model's weights based on the new dataset while retaining the acquired knowledge and contextual representations from the original training. By leveraging the pre-existing model, we are able to customise it to the needs of our storyline generation task, while minimising the required time and resources.

Model 2: AutoModel for Causal LM Model

The main model for this task will be one of the classes in the Hugging Face Transformers library, AutoModel for Causal LM. AutoModel LM is a general-purpose architecture for building various types of neural network models, including language models. It provides a convenient way to instantiate and customise pre-trained models from the Hugging Face Transformers library, such as GPT-2. As a result, we believe this as a suitable model for language generation, where the goal is to generate text with one token at a time, conditioned on the previously generated tokens to ensure generated texts' coherence and structure.

The following simplified steps were taken to fine-tune the AutoModel for Causal LM to generate our storyline (as seen in Figure 2):

1. Import AutoTokenizer and AutoModel for Causal LM, and load pre-trained model GPT-2
2. Split data into training and validation sets
3. Hyperparameter tuning
4. Define trainer with the best hyperparameters
5. Input prompt and generate the storyline
6. Evaluate

Hyperparameter Tuning

In this experiment, we will explore the baseline model by loading the pre-trained GPT-2 model from the transformers library and generate text with its pre-trained weights. For the AutoModel for Causal LM, we will fine-tune particularly on the learning rate, weight decays and epochs.

Arcane - Specific	{'learning_rate': 0.00005, 'weight_decay': 0.001, 'epoch': 2}
Expanded	{'learning_rate': 0.0001, 'weight_decay': 0.01, 'epoch': 3}

We conducted grid search on the arcane-specific and expanded training set. We identified the best hyper-parameters with the respective values that produced the lowest validation loss for each scenario in the table above.

4. Results

Our code can be found in the Google Collab links in appendix at the end of this paper**.** Below we include a table of results for both our baseline model and our fine-tuned GPT-2 model using our chosen prompts. The results will be displayed in a tabular format to compare the original with the generated texts.

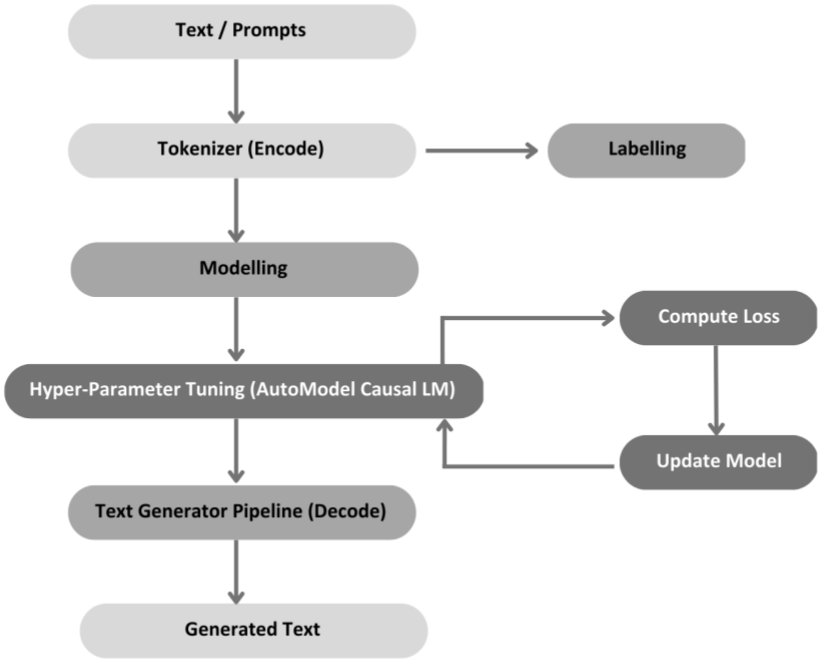


Figure 2: Computational Pipeline (Remake ourselves adapting to our project)

4.1. Baseline Results

Arcane Text	Baseline Generated Text
As dissent reaches a fever pitch, two sisters steal an artefact of untold power. Discovery and danger intertwine as heroes are born and bonds are broken. Will this power change the world, or lead it to ruin? This is the world of Arcane. Nestled at the heart of the continent, Piltover stands on the precipice of progress. The invention of hextech shines with subtle promise of a glorious future, and two brilliant scientists lead the charge. Such potential is not without cost, and failure could prove devastating as politicians vie for a power they barely understand. As unrest stirs in the Undercity and upheaval looms, the people of Piltover must protect their future… no matter the cost. In the shadows of progress, far below the ivory towers of Piltover, the Undercity waits. The air there is thick with shouts and smoke, and each alleyway promises danger, desire...or both. But at its heart, this sister city thrums with hope, ingenuity, and an unbreakable spirit that neither greedy chembarons nor aggressive enforcers could ever take away. As the age of hextech looms, her citizens no longer see themselves as Topsiders. Instead, they look up with desperate determination and dream of something more.	<u>Two sisters. Two cities. One discovery that will change the world forever. In the cities of Piltover and Zaun, unrest stirs as inventors and thieves, politicians and crime lords chafe against the constraints of a society torn asunder.</u> And in a land that should be more than just a home, the world is torn asunder. Two lives for life. One from the start. The story begins in Piapla, a town of 400,000 but also known as the "The Great River." The city began to crumble before all eyes in 1802 when the world's two main trading centres in the world met (the first world to be independent), just north of the English Channel, in the river Caledonia. And it had a history. The discovery of gold in the early centuries of modern times was only beginning to pay off. For most of the late 19th and early 20th centuries British trade vessels, and in many places even the English, were travelling through its waters every day. A couple of miles inland in the English Channel, a small town called Piapla and its harbour were a few miles of secluded harbour. But at dusk and dawn they suddenly felt the sharp,

Figure 3: Baseline Results comparing Original Text to Generated Text.

4.2. *Arcane-Specific Results/Temperature*

Arcane Text	Fine-Tuned Arcane-Specific Generated Text
As dissent reaches a fever pitch, two sisters steal an artefact of untold power. Discovery and danger intertwine as heroes are born and bonds are broken. Will this power change the world, or lead it to ruin? This is the world of Arcane. Nestled at the heart of the continent, Piltover stands on the precipice of progress. The invention of hextech shines with subtle promise of a glorious future, and two brilliant scientists lead the charge. Such potential is not without cost, and failure could prove devastating as politicians vie for a power they barely understand. As unrest stirs in the Undercity and upheaval looms, the people of Piltover must protect their future… no matter the cost. In the shadows of progress, far below the ivory towers of Piltover, the Undercity waits. The air there is thick with shouts and smoke, and each alleyway promises danger, desire...or both. But at its heart, this sister city thrums with hope, ingenuity, and an unbreakable spirit that neither greedy chembarons nor aggressive enforcers could ever take away. As the age of hextech looms, her citizens no longer see themselves as Topsiders. Instead, they look up with desperate determination and dream of something more.	Two sisters. Two cities. One discovery that will change the world forever. In the cities of Piltover and Zaun, unrest stirs as inventors and thieves, politicians and crime lords chafe against the constraints of a society torn asunder. Piltover is a city of trade and commerce that, as it is now, is the best of the great civilizations for their region. Piltover is where foreign goods can be found and sold, and where trade is made as easy as pie. Piltover is where innovation can thrive, and where innovation is always welcomed. Piltover is where the very fabric of humanity grows with each passing day, and where innovation is all that lies beneath. Piltover's future is uncertain.

Figure 4: Fine-Tuned Model (arcane-specific) Results comparing Original Text to Generated Text

4.3. *Expanded Results*

Text #	Arcane Text	Fine-Tuned Expanded Generated Text
1	Prompt 1: "Two sisters. Two cities. One discovery that will change the world forever. In the cities of Piltover and Zaun, unrest stirs as inventors and thieves, politicians and crime lords chafe against the constraints of a society torn asunder."	
	As dissent reaches a fever pitch, two sisters steal an artefact of untold power. Discovery and danger intertwine as heroes are born and bonds are broken. Will this power change the world, or lead it to ruin? This is the world of Arcane. Nestled at the heart of the continent, Piltover stands on the precipice of progress. The invention of hextech shines with subtle promise of a glorious future, and two brilliant scientists lead the charge. Such potential is not without cost, and failure could prove devastating as politicians vie for a power they barely understand. As unrest stirs in the Undercity and upheaval looms, the people of Piltover must protect their future… no matter the cost. In the shadows of progress, far below the ivory towers of Piltover, the Undercity waits. The air there is thick with shouts and smoke, and each alleyway promises danger, desire...or both. But at its heart, this sister city thrums with hope, ingenuity, and an unbreakable spirit that neither greedy chembarons nor aggressive enforcers could ever take away. As the age of hextech looms, her citizens no longer see themselves as Topsiders. Instead, they look up with desperate determination and dream of something more.	Piltover is a city of trade and commerce that, as it is now, is the best of the great civilizations for their region. Piltover is where foreign goods can be found and sold, and where trade is made as easy as pie. Piltover is where innovation can thrive, and where innovation is always welcomed. Piltover is where the very fabric of humanity grows with each passing day, and where innovation is all that lies beneath. Piltover's future is uncertain.
2	Prompt 2: "Piltover is a thriving, progressive city whose power and influence is on the rise. It is Valoran's cultural centre, where art, craftsmanship, trade and innovation walk hand in hand. Its power comes not through military might, but the engines of commerce and forward thinking."	
	Situated on the cliffs above the district of Zaun and overlooking the ocean, fleets of ships pass through its titanic sea-gates, bringing goods from all over the world. The wealth this generates has given rise to an unprecedented boom in the	Piltover is a thriving, progressive city whose power and influence is on the rise. It is Valoran's cultural centre, where art, craftsmanship, trade and innovation walk hand in hand. Its power comes not through military might, but the engines of commerce and forward thinking. -

	city's growth. Piltover has - and still is - reinventing itself as a city where fortunes can be made and dreams can be lived. Burgeoning merchant clans fund development in the most incredible endeavors: grand artistic follies, esoteric hextech research, and architectural monuments to their power. With ever more inventors delving into the emergent lore of hextech, Piltover has become a lodestone for the most skilled craftsmen the world over.	Ulla. She has spent her life as a member of the Order of the Shadow, a new alliance of guardians fighting against the darkin's encroachment on a realm ruled by powerful, but monolithic empires. In that conflict, they have defeated the enemy of all who oppose their power and been victorious. -Senna. -Aquinas. He has spent his life fighting to end the war in Kalimdor, fighting to overturn a new coalition of celestial allies that are already encroaching upon the land.
3	Prompt 3: "The empire of Shurima was once a thriving civilization that spanned an entire continent. Forged in a bygone age by the mighty god-warriors of the Ascended Host, it united all the disparate peoples of the south, and enforced a lasting peace between them."	
	Few dared to rebel. Those that did, like the accursed nation of Icathia, were crushed without mercy. However, after several thousand years of growth and prosperity, the failed Ascension of Shurima's last emperor left the capital in ruins, and tales of the empire's former glory became little more than myth. Now, most of the nomadic inhabitants of Shurima's deserts eke out a meagre existence from the unforgiving land. Some have built small outposts to defend the few oases, while others delve into long lost catacombs in search of the untold riches that must surely lay buried there. There are also those who live as mercenaries, taking coin for their service before disappearing back into the lawless wastelands. Still, a handful dare to dream of a return to the old ways. Indeed, more recently the tribes have been stirred by whispers from the heart of the desert—that their emperor Azir has returned, to lead them into a new, wondrous age.	The empire of Shurima was once a thriving civilization that spanned an entire continent. Forged in a bygone age by the mighty god-warriors of the Ascended Host, it united all the disparate peoples of the south, and enforced a lasting peace between them. There are many tales of how he came to be at the centre of this conflict, and his first appearance was when a band of nomadic warriors invaded the neighbouring land. It was during the war that Nasus's first memories came to be. At the height of it all, he was one of the few survivors of an ancient battle, and endured the bitter end to it all. The Shuriman warbands of Noxus had marched onto the eastern coast of the Noxus to ravage the lands and villages of Runeterra.

Figure 5: Fine-Tuned Model (Expanded) Results comparing Original Text to Generated Text

4.4. Final Result Analysis

Our models were able to produce results that were linguistically and grammatically coherent, as well as contextually relevant (Figures 4 and 5).

The **baseline** and **arcane-specific** models were compared with the same prompt and original text, whereas the **expanded** model was evaluated via three different prompts. Details on results and model evaluation will be further discussed in the next section.

5. Evaluation

Generating text is a complex domain, and academia and the industry face difficulties in determining the best measures for determining the effectiveness of generative models. Each generative task is unique and possesses its own intricacies and distinctiveness; for instance, metrics used to assess dialogue systems differ from those used for summarization. While metrics such as perplexity, BLEU/ROUGE and other metrics can provide a quantitative measure of text generation models' performance, they have limitations in fully capturing the quality of generated text (Kawthekar et al., 2017). For the purpose of our project, due to the inherently subjective nature of text generation quality (particularly in the context of game storyline generation, which may contain specialised vocabulary specific to the game's lore), we decided to incorporate human evaluations in our assessment of the model's performance. This can be used to provide a more comprehensive understanding of the model's effectiveness as it provides a subjective assessment of factors such as fluency, coherence, and relevance. Incorporating human evaluation alongside metrics is important to gain a complete understanding of how well text generation models are performing and to identify areas for improvement that may not be captured by the metrics alone.

Model Evaluation Metrics

5.1. Validation Loss: Tuned model

In our text generation project, we fine-tuned two AutoModel Language Models, one based on the Arcane text model and the other based on all the champion descriptions in League of Legends. The best hyperparameter configuration for the Arcane-based model, yielding an evaluation loss of 1.278050. Meanwhile, the best hyperparameter configuration for the model based on champion descriptions, resulting in an evaluation loss of 1.2700029611587524. These optimised models demonstrate reduced validation losses compared to the GPT-2 baseline model, indicating improved performance in generating relevant text.

5.2. Perplexity: Text Generation Evaluation

5.2.1. Cosine Similarity with Arcane Description

To further evaluate the text generation performance of our fine-tuned models, we utilised perplexity and BLEU scores. Perplexity measures how well a model predicts a sample, with lower values indicating better performance. In our case, the Arcane-based model achieved a perplexity of 3.5896338, while the champion description-based model had a perplexity of 3.560863. These results suggest that both models can generate reasonably coherent and contextually relevant text.

Additionally, we calculated the BLEU score for each model, which assesses the similarity between the generated text and the original descriptions. The Arcane-based model achieved a BLEU score of 0.0028014, while the all-champion description-based model achieved a slightly higher score of 0.0033288. These scores show that the generated text is not very similar to the original descriptions, but our fine-tuned models still produce text that is related to the context.

Our low BLEU score can be explained by the following:

1. Inherent limitations of BLEU: The BLEU metric does not account for other aspects of language quality, such as coherence, grammatical correctness, or semantic richness; therefore, a low BLEU score does not necessarily mean the model is bad.
2. Diverse and creative output: For the model generating diverse and creative text that deviates from the original reference can result in low BLEU scores.
3. The mismatch between reference and generated data: The low BLEU score might also be due to a mismatch between the generated text and the reference data.

5.3. Similarity Analysis (Model Accuracy/Performance): BERT Model and Human Evaluation

5.3.1. Human Evaluation

We conducted an online survey where 20 participants were each presented with 6 pieces of text (3 original and the corresponding 3 generated by our fine-tuned model), ordered randomly. Each participant was asked to rank the texts on a scale from 1-10 (10 being best) for the following properties in response to the questions in the table on the right. Results from the survey can be found in figures 6 and 7.

Language quality	'On a scale from 1-10, how well does the text below make use of proper and correct English?'
Coherence	'On a scale from 1-10, how well does the text below flow in terms of storyline coherence?'
Creativity	'On a scale from 1-10, how much do you agree with the following statements?'

6. Experiment Limitations

6.1. Dataset Limitation

The amount of data will not be considered a limitation as our task is to fine-tune a pre-trained model to our smaller and specific storyline. We identified region classification and character relationships as limitations to our dataset. These are crucial to our model training, as we do not wish the generated storylines to have region overlaps or incorrect champion relationships that do not make sense.

Regional classification: regions for each champion are vaguely mentioned in the character descriptions, thus, the model cannot classify where each champion comes from.

Character relationships: character relationships remain unclear throughout the descriptions. For example, Garen is Lux's older brother, but the model could interpret them as friends.

6.2. Model Limitations

While GPT-2 remains as a highly advanced and impressive model in the field of NLP, there are several limitations that must be taken into account:

Black box. In any deep learning algorithm, the process behind the training and predictions remain blurred, such that models like GPT-2 lack transparency and interpretability that underlies the algorithm.

Potential bias. Eventhough we fine-tuned the model with our text data, the model is still prone to capturing bias/noise from its pre-trained large corpus of text.

Lack of common sense. The GPT-2 model may not have the ability to understand contextual meaning like humans do. This is associated with the dataset limitations, where the model's inability to classify regions and understand character relationships through the descriptions provided.

Dependence on input quality. The input is as important as the output. Alike ChatGPT, high-quality detailed prompts generate high-quality texts, and vice-versa. The concept of garbage-in-garbage-out applies to our project task.

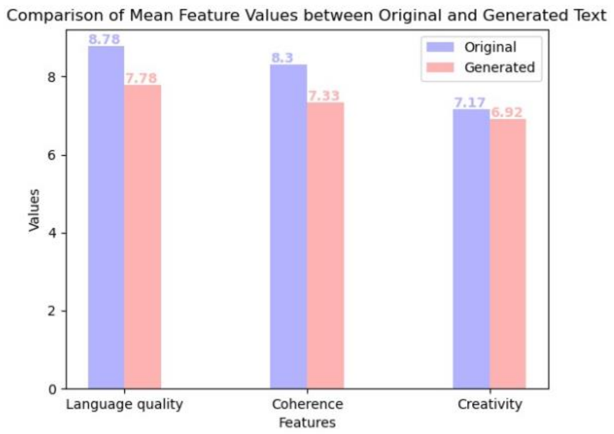


Figure 6: Average rating across the 3 properties for the original and generated text. We collected ratings from 20 participants. Each participant rated 3 generated texts and 3 original texts on each property indicated above.

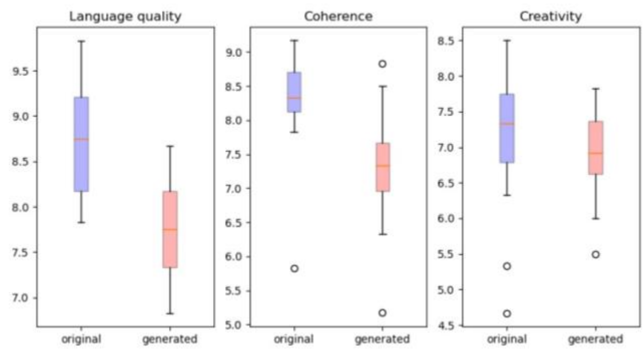


Figure 7: Distribution of ratings of each property for original and generated texts.

6.3. Computational limitation

The nature of building/fine-tuning a state-of-art language model requires massive computational power. Therefore, GPT-4 and GPT-3 models were the optimal choices regarding output quality. However, due to the computational and time constraints, we chose to experiment with GPT-2, which contains fewer parameters but is pre-trained on a smaller corpus of text.

7. Conclusion & Discussion

Our study found that by using unique and context-specific tokens to teach GPT-2 a specific output structure by delimiting the training set and prompts was a successful approach for text generation in the gaming context. Although the fine-tuned model scored worse on all three properties, the generated outputs were on average, on par with human-written dialogues, and the ratings distributions showed that GPT-2 could learn the linguistic style and structure of League of Legends texts. One advantage of the fine-tuned model is that it can create large numbers of generated storyline descriptions from the same prompt, which is fast and low-costly. As a feasible alternative to writing new storylines by hand, modifying the most creative outputs or cherry-picking best quality outputs can be a viable alternative. Additionally, to improve output quality, further studies to search for the optimal generation temperature can be conducted.

A limitation of our project however, is that the fine-tuned GPT-2 model does not generalise yet to different role-playing games, given it is context specific to League of Legends. This is because the resulting outputs contain references to the lore of League of Legends, meaning it is not easily transferable to other game worlds. To solve this problem, several approaches can be taken to make the model more generalise better to a broader range of games. Firstly, substituting all named entities by a [NAME], [LOCATION], or [FACTION] tag before the fine-tuning process could make the outputs more generic. Then, the placeholders can be filled with names, locations or lore from a different game world, either automatically using a language model, or by hand.

As an alternative approach, we could also explore the potential of translating outputs from one game to another, through machine translation techniques in a post-processing step. For example, using distributional semantics to translate game-specific terminology. This would entail building a vector space of word embeddings using words retrieved from two different games. If it turns out that certain words from different games are often used in the same context, these words would appear closer to eachother in the resulting vector space, and we could use this information to substitute one word for another in generated outputs.

For future works, to improve the model's ability to generalise to several games, we could also fine-tune GPT-2 on a heterogeneous dataset from different video games. The dataset would contain additional tags that displays the game world or game genre that a specific datapoint originates from. These tags could then be used to direct the model towards outputs for the specific genre or game, assuming the dataset is varied enough to capture significant differences in style and content between games.

In addition, future studies can investigate the potential of generating storyline titles and objectives given a particular dialogue, which can be seen as a form of text summarisation. This is an interesting field given that text coherence is a result of the link between storyline connection to the dialogue. Moreover, evaluating the impact of various temperature settings to observe their effects could also be insightful, although it may not be straightforward given that these findings may not easily generalise to models trained on different game data.

To exercise more control over generation, we should explore whether we can add additional annotation tags to the training set, such as capturing expressed sentiment and NPC-player relations, which could be used in prompts to guide the generator towards outputs with desired properties. Using the largest GPT model (GPT-4) might also improve the language quality of the generated examples. However, if we start using larger pre-trained language models, we must also investigate whether the size of the training set should be increased proportionately. It would also be interesting to find out how large a dataset of game texts should be before it can be used to teach GPT-2 the structure and linguistic style of game texts.

Finally, it is worth noting that culture references may exist in the generated dialogue, since GPT-2 was pre-trained on web text, therefore if we fine-tune GPT-2 on a dataset of video game text from multiple games, the amount of references to other games will also increase, which could be perceived as a fun Easter-eggs by players.

8. References

Ammanabrolu, P. *et al.* (2019) "Guided Neural Language Generation for automated storytelling," *Proceedings of the Second Workshop on Storytelling* [Preprint]. Available at: https://doi.org/10.18653/v1/w19-3405.

arcane.com. (n.d.). *Arcane: League of Legends*. [online] Available at: https://arcane.com/en-gb/.

Boden, M.A. (2007) "Creativity in a Nutshell," *Think*, 5(15), pp. 83–96. Available at: https://doi.org/10.1017/s147717560000230x.

Burtenshaw, B. (2018) "A brief introduction to natural language generation within computational creativity," *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)* [Preprint]. Available at: https://doi.org/10.18653/v1/w18-6601.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805.

Openai, A., Openai, K., Openai, T. and Openai, I. (2018). *Improving Language Understanding by Generative Pre-Training*. [online] Available at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. [online] Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Schuff, H., Vanderlyn, L., Adel, H. and Vu, N.T. (2023). How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, [online] pp.1–24. doi:https://doi.org/10.1017/S1351324922000535.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan N, Kaiser, L. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762.

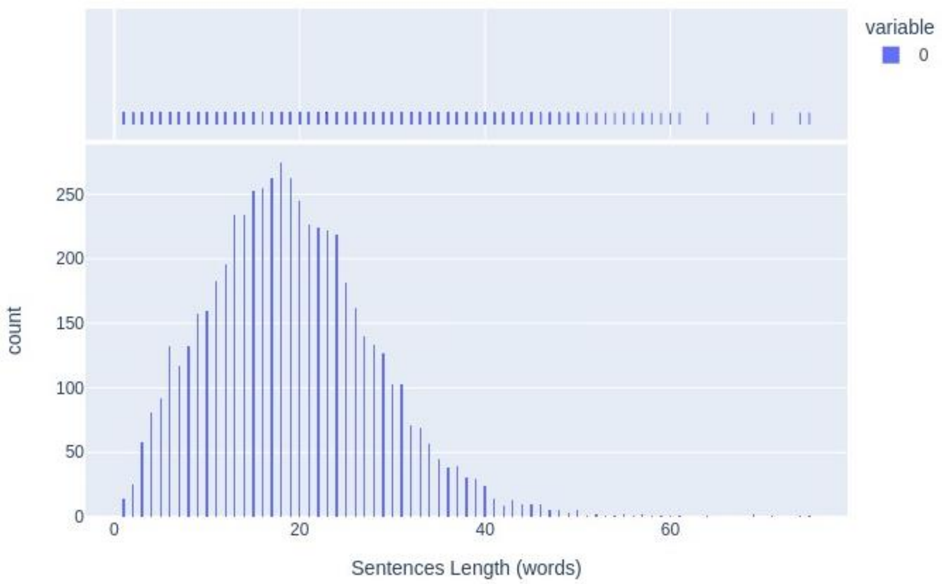
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Brew, J. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. [online] Available at: https://arxiv.org/abs/1910.03771.

www.leagueoflegends.com. (n.d.). *League of Legends*. [online] Available at: https://www.leagueoflegends.com/en-gb/.

9. Appendix

Google Collab Link to Arcane Model:
https://colab.research.google.com/drive/17jsJZHjFsZlRaltbS1ohmVemB7ZejF_M?usp=share_link

Google Collab Link to our Full Model:
https://colab.research.google.com/drive/1RPtHuims_SuNXS3NoxQDXq-ZCTEF_-mK?usp=share_link



Appendix 1: Distribution of sentence length (in words)

Most frequent words in all the biography

Appendix 2: Most frequent words in our vocabulary