Title Generation for Fictional Stories

Jacob Malin and **Tony Diep** and **Oscar Wiestling** and **Cody Cayetano**
Department of Computer Science, University of Minnesota-Twin Cities
{malin146,diep0015,wiest026,cayet010}@umn.edu

1 Abstract

In this paper we plan to explore the creative potential of large language models, focusing on fine-tuning a GPT-2 model for generating titles for fictional short stories, while contrasting the results to ChatGPT.(OpenAI., 2023b) Our motivation is to evaluate large language models on a creative generation task while avoiding traditional summarization-based approaches. The dataset was acquired from Hugging Face and is made up of horror stories gathered from Reddit. Human evaluations show that our fine-tuned GPT-2 model often surpassed ChatGPT in creativity, while our blind testing preferred ChatGPT for what people expected to see as titles in Reddit. This study showcases the very promising ability of large language models to be able to perform automated creative content generation.

2 Introduction

2.1 Motivation

The motivation behind this project is to assess the capacity of a language model to tackle more subjective and abstract tasks. A top-level survey of the literature reveals that the concept of title generation has primarily been applied to factual writing, such as news articles and academic papers. In such cases, generating a title represents a relatively objective task, as it relies on summarizing the content or extracting an important quote from the text. While acknowledging the effectiveness of these approaches, our project aims to explore how well a language model can generate titles for fictional works and how effectively it can capture the subjective elements of a story, such as its tone and themes.

2.2 Objective

The end goal is to compare the GPT-2 generated titles for fictional pieces against their original titles, hopefully showing that the language model has the ability to handle more creative and nuanced aspects of language generation, and can be used for more than just summarization. On top of this goal, the fine-tuned GPT-2 model should be more creative and varied than the titles generated by GPT-3.

2.3 Literature Survey

Other researchers have looked into the idea of generating titles for various types of written works, and while these have been successful, they generally focus on summarizing key points of the writing rather than trying to generate something creative based on it. The following is a brief summarization of three different works focusing on title generation:

The reviewed literature encompasses three research papers, each addressing distinct facets of text generation. In "TCR: Short Video Title Generation and Cover Selection with Attention Refinement," the authors introduce a novel approach utilizing a transformer model with a refinement module to jointly generate creative short video titles and select covers.(Yu, 2023) Their method, focusing on the limitations of existing language models in generating titles for entertaining content, incorporates a new dataset for training and evaluation, leveraging both automated metrics and human judgment.

Shifting the focus to document title generation, "A New Probabilistic Model for Title Generation" proposes an innovative probabilistic model with an 'information source' step to improve title quality.(Rong Jin, 2002) Outperforming traditional models in both F1 score and human judgments, the research highlights the crucial role of automatic title generation in swiftly conveying a document's main idea. The model effectively addresses issues such as overused common words by introducing the concept of "commonness," demonstrating its success through quantitative measures and human evaluations.

Meanwhile, "WriterForcing: Generating more

interesting story endings" explores enhancing creativity in text generation.(Gupta, 2019) The authors introduce a model that prioritizes important keyphrases using the ITF loss function, leveraging the RAKE algorithm to identify them. While excelling in introducing uncommon tokens and proper nouns, the model occasionally deviates from the original story context, showcasing both strengths and limitations in generating diverse and interesting text. Together, these papers contribute insights into evolutions in text generation across various domains.

The surveyed research faces challenges such as the subjective nature of evaluating creative text generation, with the need for human judgment being crucial because it is very hard to automatically gauge the "creativeness" of a written work. Additionally, there are limitations in ensuring that generated content aligns closely with the original context.

2.4 Impact

People who are interested in the limits of AI will be interested in this experiment. How creative can AI be? Auto-generated titles can also help people create titles for the content they write. Forum users of sites like Reddit and Twitter can quickly create creative titles for the things they wrote to get more attention. Generated creative titles may also inspire writers who have writer's block. However, some negative consequences are taking away more "human" involvement in creating content. Auto-generated creative content has already been criticized in the past for its lack of originality since LLMs are trained on other people's data.

3 Approach

3.1 Hypothesis

We hypothesized that fine-tuning GPT-2 to generate creative titles for short stories would outperform ChatGPT based on human evaluation of creativity.

3.2 Methods

We used a finetuned GPT-2 model to generate stories from the intone/horror_stores_reddit database on Hugging Face. We then compared that to few-shot prompting with ChatGPT.

3.2.1 GPT-2 Fine-tuning

Our methods utilized a pre-trained GPT-2 model from the popular machine learning website, Huggingface, in which we finetuned the model for the task of title generation. Due to resource availability we were limited to a version of GPT-2 with 124 Million hyperparameters, thus the performance on the task could not be compared across models of larger size such as GPT-2-Large. With the language model selected we found an appropriate dataset for the task, intone/horror_stories_reddit. (intone, 2023) The dataset is composed of 5.61k rows of short-form horror stories with columns like 'title', 'text', 'url', and 'subreddit'. For our purposes, we used the 'title' and 'text' columns in our data points and the 'subreddit' column for filtering. The dataset was very noisy including hyperlinks, emojis, and long strings of the underscore character used as visual separators. To build our data points we first filtered through the stories either by the 'subreddit' column or the word count of the 'text' column. When filtering by word count it was important to enforce a MIN_WORDS limit to disclude any very short stories such as the stories belonging to the 'TwoSentenceHorror' subreddit, and a MAX_WORDS limit to not exceed the max_sequence_length = 1024 of the model. After filtering the rows we denoised the 'text' and 'title' columns before building the datapoints. The data points were constructed using the following formula:

$$\text{`}text\text{'} + trigger + \text{`}title\text{'}$$

where trigger is substituted with the string "Title:". After constructing the data points they were tokenized and passed to the trainer for finetuning. The model was finetuned with a train and eval batch_size = 1, due to memory limits, and a learning_rate = 2e-5 across 4 epochs. After finetuning, the model was prompted to generate titles using the following input: 'text' + trigger. The titles were generated using greedy, beam, top-k, and top-p decoding methods.

3.2.2 Prompting ChatGPT To Create More Creative Titles

ChatGPT would often produce very formulaic and generic titles when using zero-shot prompting to generate titles for stories. (OpenAI., 2023a) It often had a verb+adj+noun combination of words that felt inorganic, followed by a colon and a quick sentence summary. An example of this would be: "Shadows of Change: A Father's Awakening to the Consequences of Energy Choices". To improve ChatGPT output, we used a few-shot prompting where we supplied the model with the top 10 rated

titles on Goodreads and a 4-5 word limit for the titles. An improved title to the previous example was: "Shadows in the Energy Mist". We also discovered that if we gave the model too many details and instructions, it would result in the model not following all the instructions and forgetting many details. It was also advantageous to input the story first before giving the instructions to follow.

3.3 Challenges

We encountered a lot of computing difficulties. Ranging from not enough RAM to the computation time itself. For both of these models, we also struggled with the input token length. This limited us to very short-form stories that we could generate titles for. Because we could only use short-form stories, it was also difficult to find a database whose stories were around 600 words or less. Therefore we had to sift through databases to find stories that fit our restrictions. For GPT-2, we failed to make the titles sound less like a summary. And for ChatGPT, we couldn't find a way for it to stop using very generic words.

4 Results

4.1 Experimental Refinements

Seeking to enhance our results, we experimented with two different approaches. Surprisingly, despite our efforts, these approaches did not yield the expected refinement in our results. The following sections detail each approach, explaining their methods, rationale, and comparative results.

4.1.1 Finetuning a Different Model

Since one of the more significant limitations in our finetuning of the GPT-2 model was its max_sequence_length = 1024, forcing us to enforce a MAX_WORDS limit on the 'text' component of our datapoints, we searched for a model with a larger max_sequence_length. The first intuition was to use GPT-3 but the model was not available on HuggingFace. Instead, we opted for GPT-Neo, a model based on the GPT-3 architecture with a max_sequence_length = 2048.(Eluther AI., 2023) With the larger max_sequence_length, we were able to finetune with longer stories consequently increasing the amount of training data from 300 with GPT-2 to (will look later) with GPT-Neo. Despite the increased amount of training data and the use of a more modern architecture, GPT-Neo generated poorer results compared to GPT-2.

GPT-2	**GPT-Neo**
The Dream House	The house in front of us
What is this scary music box instrumental playing in my house	Halloween Lofi Music
The Night I Wasnt Alone	I was scared to go to sleep

Table 1: Comparing outputs between GPT-2 and GPT-Neo

This could be due to a couple of factors. Firstly, the complexity of the model's architecture might have introduced challenges affecting its ability to generalize effectively. Additionally, potential overfitting due to the larger training data could have affected the generated results.

4.1.2 Utilizing a Refinement Module

Taking inspiration from the paper "Automatic Title Generation for Text with Pre-trained Transformer Language Model" we built a refinement module to possibly enhance the finetuned GPT-2 model generated titles by pushing them towards their respective ground truth titles. (Mishra, 2021) Following similar methods to Mishra et.al, we finetuned a separate GPT-2 model on a dataset composed of our generated titles and the ground truth titles. This dataset was built by first filtering out a subreddit 'scarystories' in our training data that was used on our finetuned GPT-2 model used for title generation to ensure the title generation model was not finetuned on this data. We then generated titles for each datapoint built from the 'scarystories' subreddit. Since we generated titles using four different decoding methods, and the use of the dataset is to push a generated title more towards the ground truth title, choosing the appropriate title from the four generated titles was an important step.

To choose the appropriate title we compared the ROUGE scores between the four generated titles and the ground truth title, choosing the title with the highest score above a minimum ROUGE score, to eliminate titles too far from the ground truth. With the selected titles we built the datapoints using the following formula:

$$`pred - title' + `gt - title' + trigger$$

Where *'pred-title'* is our generated title, *'gt-title'* is the ground truth title, and *trigger* is substituted

with the string “Title:”. We then fine-tuned the refinement model using the same training arguments as the title generation model on the new dataset.

Surprisingly the text generated through the refinement module was less like titles generated by the generative model as they repeated words and in some cases would resemble the text from the story. One potential reason for this might be that the model used for refinement was not able to learn much context from the newly constructed data points composed of the generated titles and their corresponding ground truth titles. To potentially address this issue we adjusted the construction of the data points to use the formula:

$$‘text’ + sep + ‘pred - title’ + ‘gt - title’ + trigger$$

Where *sep* is substituted with the string ‘Start-Pred:’ and *trigger* remains the same. Fine-tuning the refinement module on the new data points produced similar, non-title-resembling, results. Future work on adjusting the loss function for fine-tuning the refinement module might produce better results.

4.2 Human Evaluation

4.2.1 Measuring Creativity

To measure the success of the models for generating titles, we used humans to evaluate the titles based on their creativity. We defined creativity based on 3 attributes defined in a paper by Franceschelli and Musolesi (2023):

1. Is the title novel or new?
2. Is the title surprising?
3. Is the title valuable? Does it provide insight and interest into the story?

With these 3 ideas defined, we measured the generated titles on a score of [1,5] per attribute and then added those scores together on a scale of [3,15] and averaged them between evaluators, and then finally, chose which of the three titles rated the highest.

For example for the story, "The stalker", ChatGPT generated "Haunted Hotel Horror" and our model generated "The Night I Wasnt Alone". We then rated them as can be seen in Figure 1.

Overall, GPT-2 was selected 55.6% of the time, the most picked out of the 3 options. The novelty and surprise criteria affected our choice the most. Although ChatGPT produced titles that had high value, meaning that it provided context to the story, it often failed in the novelty and surprise criteria. As seen in Figure 5, ChatGPT often reused words

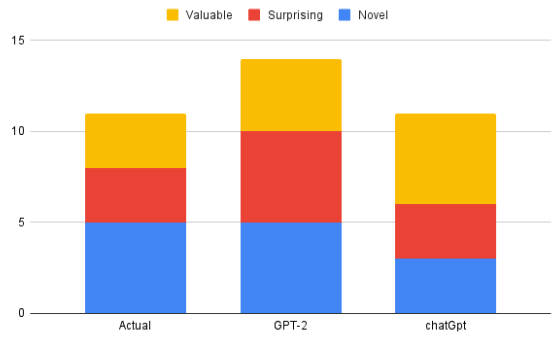

Figure 1: An example rating for "The stalker"

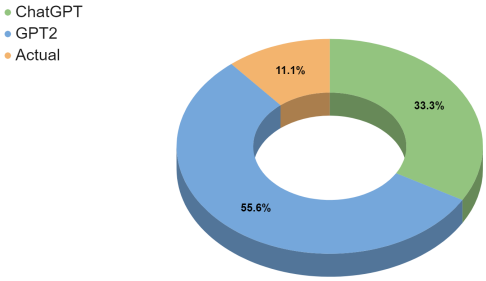

Figure 2: Creativity Evaluation Pick Rate

for the titles it generated. While GPT-2 generated titles that were often unique for its story in comparison to other stories. As a result, GPT-2 scored higher than ChatGPT in our human evaluation.

4.2.2 Blind Survey

Recognizing the presence of bias in these results we also gathered an unbiased preference of titles via a blind test conducted at the poster presentation. The methods of the blind test experiment were designed as follows:

1. Place three unlabeled titles for the same story next to each other. (The labels would be: ‘Actual’, ‘Fine Tuned GPT-2’, ‘ChatGPT’)
2. Direct class members to pick the preferred title using prompts like “Choose the title that you find most horrifying or that would make you want to read the story”.
3. Count the number of times class members chose titles of the aforementioned labels.

The results from the blind survey that we did during the poster presentation can be broken down in two ways, per participant and per story. There were 15 participants and 10 stories in which each participant chose their favorite among the 3 provided titles. Overall ChatGPT had the highest selection rate, per person and per story. Our model was not

too far behind, and the actual title did not perform very well. The actual title never won per person, and only won once per story. The crowd favorites from our model were “The Night I Wasnt Alone” and “I’m not sure if I should post this”. Our model won over the actual title 73% of the time per person, and 60% per story.

One interesting takeaway from our presentation survey is that AI could easily replace creative subreddit title generation with great success. Overall the actual responses performed very poorly, and often when people were taking the survey, they would guess that our model was the actual responses, due to how simple and boring many of the actual titles were, and how interesting our titles were in comparison. People did not often mistake the ChatGPT title for the actual titles, as ChatGPT has a very recognizable and very distinct style. The titles that performed the best from our model were also the most intriguing overall, those that leave the user in suspense. The best titles from ChatGPT were ones with strong adjectives, which while very distinctive of ChatGPT, still had wide popularity.

4.2.3 Word Usage

Below is a visual representation of the frequency of words used in generating titles for scary stories pulled from Reddit data. Where frequently used words are larger than infrequent words. It can be seen that the GPT-2 Titles frequently used the word “house”, this is similar to the actual titles of the story shown in the next figure. This data shows that GPT-2 titles may be more likely to use a title that is summary-like because the title includes words frequently found in the stories.

An example of this is the word house, many of the authors wrote stories that involve their homes, therefore GPT-2 titles often included the word “house”. When we compare it to the actual titles, we can see that authors also often use the word “house” in their titles. Besides that, we can see that the other words used in the GPT-2 titles are infrequently used. Meaning that it often produced new, novel, dissimilar titles for different stories. On the other hand, ChatGPT has many large words in its figure. This means that ChatGPT often reuses words for different stories. An example of this is the word “shadow”. This data shows how ChatGPT often only summarizes and understands the story on a baseline level, which causes it to generate very generic and boring titles even for scary stories that may have different plots.

Figure 3: Fine-tuned GPT-2 Titles

Figure 4: Actual Titles

Figure 5: ChatGPT Titles

5 Discussion

5.1 Reproducibility

The human evaluation is reproducible, but the data generated from the poster session is most likely not as reproducible. The data we got with our human evaluation has consistent standards for what defines creativity, and that data was then aggregated to get a single score for all annotated titles. This means that it would be easy to redo. However, it should be noted that since we were the ones to do the human annotations, a very good idea of what qualities of creativity that each model was producing, so there is most likely some bias in our final results.

On the other hand, while the data from the poster session is very interesting, and it gives a very good insight into the opinions of a wider audience, it is not easily reproducible. This is because we did not use a set script while having users fill out our survey, and we also talked about the differences between ChatGPT versus our model, which could have influenced the results greatly. Were we to do a more formal study, we would most likely have a set script to instruct the user on their task, and also probably give the original story to the user so that they would be able to choose the title based on how well it matches the story.

5.2 Databases

Due to the limitations of our model, we were unable to use longer stories, so we restricted the stories that we used with our model to only those that were less than 600 words. We also used a random sample from the dataset for our human annotations.

5.3 Ethics

There are many ethical considerations for this as these are both forum posts that were not collected with the direct consent of the users of the forum, and also the titles that we are training our model on are creative works, which has similar ethical concerns to generating art using AI tools. Even if we retrained our model on a new dataset that was collected solely for the purpose of our model, there are still concerns about how our model could be used by other people. For example, someone could use short creative generation like this to automatically replace titles to make more people click on forum posts, which would negatively impact authors' control over their own work.

5.4 Limitations/Future Research

As stated above, a major limitation of our model was that there was a token limitation set by GPT-2 which prevented us from using the entire dataset that we got from hugging face. If we were able to use the whole dataset, it is possible that we could have even more diverse/creative output from our model. A limitation of our baseline data was that we used few-shot prompting for ChatGPT, but that did not give us a lot of control over ChatGPT to make it more similar to the Reddit horror stories, which made ChatGPT produce very generic results.

Due to time constraints, we chose to generate titles that would be applicable for any horror story, and from any subreddit. However, future research in this area could specialize titles for given genres or subreddits and attempt to generalize the concept of genre between horror stories so that it could all be done using the same model. Such as analyzing the titles from a subreddit and classifying the titles as more like clickbait or more vague, and then passing that classification along with a story to receive a title that matches the naming theme of that genre/subreddit.

References

Eluther AI. 2023. *GPT-Neo*. Large Language Model, https://huggingface.co/EleutherAI/gpt-neo-125m.

Prakhar Gupta. 2019. *WriterForcing: Generating more interesting story endings*. arXiv, Pittsburgh, PA.

intone. 2023. *Reddit Horror StoriesDataset*. HuggingFace.

Prakhr Mishra. 2021. *Automatic Title Generation for Text with Pre-trained Transformer Language Model*. IEEE, Laguna Hills, CA.

OpenAI. 2023a. *ChatGPT*. Large Language Model, https://chat.openai.com/.

OpenAI. 2023b. *GPT-2*. Large Language Model, https://huggingface.co/gpt2.

Alexander G. Hauptmann Rong Jin. 2002. *A New Probabilistic Model for Title Generation*. aclanthology, Pittsburgh, PA.

Yakun Yu. 2023. *TCR: Short Video Title Generation and Cover Selection with Attention Refinement*. arXiv, University of Alberta, Edmonton, AB, Canada.