

# “Kurosawa”: A Script Writer’s Assistant

Prerak Gandhi\*, Vishal Pramanik\*, Pushpak Bhattacharyya

Department of Computer Science and Engineering  
Indian Institute of Technology Bombay, Mumbai  
{prerakgandhi, vishalpramanik, pb}@cse.iitb.ac.in

## Abstract

Storytelling is the lifeline of the entertainment industry- movies, TV shows, and stand-up comedies, all need stories. A good and gripping script is the lifeline of storytelling and demands creativity and resource investment. Good scriptwriters are rare to find and often work under severe time pressure. Consequently, entertainment media are actively looking for automation. In this paper, we present an AI-based script-writing workbench called KUROSAWA which addresses the tasks of plot generation and script generation. Plot generation aims to generate a coherent and creative plot (600–800 words) given a prompt (15–40 words). Script generation, on the other hand, generates a scene (200–500 words) in a screenplay format from a brief description (15–40 words). Kurosawa needs data to train. We use a 4-act structure of storytelling to annotate the plot dataset manually. We create a dataset of 1000 manually annotated plots and their corresponding prompts/storylines and a gold-standard dataset of 1000 scenes with four main elements — scene headings, action lines, dialogues, and character names — tagged individually. We fine-tune GPT-3 with the above datasets to generate plots and scenes. These plots and scenes are first evaluated and then used by the scriptwriters of a large and famous media platform ErosNow<sup>1</sup>. We release the annotated datasets and the models trained on these datasets as a working benchmark for automatic movie plot and script generation.

## 1 Introduction

Movies are one of the most popular sources of entertainment for people worldwide and can be a strong medium for education and social awareness. The impact and influence of film industries can be gauged from the fact that Hollywood movies invest

100s of millions of dollars and often make box-office collections of billions of dollars. The first motion picture *The Great Train Robbery, 1903*— black & white with no sound— was created at the beginning of the 20th century. Since then, the art has gone through several transformations, and now people can instantly access 4K HD movies of their liking on any smart device.

Throughout the history of film, two of the contributors to a film’s blockbuster success have been the quality of its plot and the manner of storytelling. The appeal of the movie decreases drastically if the viewers find the plot drably predictable. Writing a creative and exciting script is, therefore, a critical necessity and is extremely challenging. Add to this the constraints of time and budget, and the need for (at least partial) automation in script writing becomes obvious.

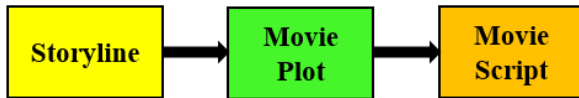
AI-based story generation has been used before. Based on the engagement-reflection cognitive explanation of writing, the computer model MEXICA (Pérez and Sharples, 2001) generates frameworks for short tales. BRUTUS (Bringsjord and Ferrucci, 1999) creates short stories with predetermined themes like treachery. With the arrival of pre-trained transformer models, automatic story generation has got a shot in the arm. Transformer models like GPT-2 and GPT-3 are extensively used for text generation. These models have shown the capability of generating creative text, albeit sometimes with hallucinations (Zhao et al., 2020). Text generated by these models also sometimes lacks coherence and cohesiveness. On the other hand, template-based models can generate coherent text but lack creativity in generating new characters and events in the plot (Kale and Rastogi, 2020).

The process of creating a movie generally starts with an idea which is then used to create a plot which is used as the base to build the movie script (Figure 1).

Novel datasets are an important feature of this

\*These authors contributed equally to this work

<sup>1</sup><https://erosnow.com/>



**Figure 1:** The thought process a scriptwriter follows in creating a movie script. An idea (**storyline**) leads to a **plot** which is then converted into a **movie script**.

paper. We closely studied the plots and prompts of movies from Bollywood and Hollywood. Such plots and prompts were scraped from Wikipedia<sup>2</sup> and IMDb<sup>3</sup>, respectively. The plots are then annotated using the 4-act story structure- an extension of the well-known 3-act structure (Field, 1979). The 4-act structure and the annotation methods are explained in detail in **appendix A.5** and **section 4**, respectively.

We introduce a dataset of 1000 Hollywood movie scenes and their short descriptions. The scripts are scraped from IMSDb<sup>4</sup>. The scenes are annotated with the four major components of a screenplay: *sluglines*, *action lines*, *character names* and *dialogues*, described in details in appendix A.4

We introduce a workbench which we call “Kurosawa”, consisting of datasets and a pair of GPT-3 (Brown et al., 2020) models fine-tuned with the said datasets. One GPT-3 model generates a movie plot given a short description of the storyline (15–40 words), while the other creates a scene based on a short description of the required scene.

Importantly, we have provided the “Kurosawa” platform to one of the biggest media platforms engaged in the business of making movies and TV shows, producing music and soundtrack etc.- to help script and content writers from different film industries create new movie plots.

#### **Our contributions in this work are as follows:**

- To the best of our knowledge, this is the first work on generating movie scenes from a scene description.
- We create and publicly release two datasets: (a) a parallel dataset of 1000 movie storylines and their corresponding plots, (b) a parallel dataset of 1000 movie scenes and their corresponding descriptions. In (a), we link available movie storylines from IMDb with available corresponding movie plots from

Wikipedia. In (b), we link available movie scenes from IMSDb with corresponding descriptions from IMDb.

- We manually annotate movie plots according to a 4-act structure which is an extension of the well-known 3-act structure (Field, 1979). Professional scriptwriters from the media and entertainment industry guided us very closely.
- We manually annotate movie scenes with four major components of a scene: *sluglines*, *action lines*, *character names* and *dialogues*, along with a short description of the scene.
- We introduce “Kurosawa”: a workbench that consists of multiple datasets and models which can assist script and scene writers in the film industry.

## **2 Motivation**

Movies are a form of visual media and can have a huge influence on life and society. Movie scripts are often 30,000 words long, comparable to a 100-page book. Though scripts can be diverse, they have fixed and oft-repeated structures, e.g., *scene heading*, *transition*, *character name*, etc.. This fixity and repetition can be dull and time-consuming and can be handed over to AI. However, a surprising fact is that AI-based models can be creative in generating novel characters and stories. These reasons have motivated the film industry to seriously consider harnessing AI for various aspects of movie making, script and scene writing being one of them.

Los Angeles Times, 19 December 2022, asks, "AI is here, and it's making movies. Is Hollywood ready?". The newspaper edition reports mainly movie editing efforts ongoing at various places using AI. Our task in the paper is allied but different in the sense that we aim to provide a "script-writers' assistant".

## **3 Related Work**

### **3.1 Automatic Story Generation**

Neural models have been able to produce stories by conditioning on different contents like visuals (Huang et al., 2016) and succinct text descriptions (Jain et al., 2017). Work on plot controllable, plan-driven story generation abounds (Riedl and Young, 2010; Fan et al., 2019; Pérez and Sharples, 2001; Rashkin et al., 2020). A related kind of work is automatic poetry generation based on keywords or descriptions (Yan, 2016; Wang et al., 2016).

<sup>2</sup><https://www.wikipedia.org/>

<sup>3</sup><https://www.imdb.com/>

<sup>4</sup><https://www.imsdb.com/>

### 3.2 Plot Generation

Plot Machines (Rashkin et al., 2020) generate multi-paragraph stories based on some outline phrases. Fan et al. (2018) introduce a hierarchical sequence-to-sequence fusion model to generate a premise and condition that in turn generate stories of up to 1000 words. This work—unlike ours—is non-neural and template-driven and is, therefore, much less creative and novel compared to what we generate.

### 3.3 Scene Generation

Automatic scene or script generation has received comparatively less attention. Dialogue generation (Li et al., 2016; Huang et al., 2018; Tang et al., 2019; Wu et al., 2019) with a semblance of scene generation has been done. There has recently been some work focusing on guiding dialogues with the help of a narrative (Zhu et al., 2020). We generate scenes in which the main elements come from a small prompt as input.

## 4 Dataset

For movie plot generation, we have taken the plots from Wikipedia. The prompts for this task have been taken from IMDb. In IMDb, this prompt can be of two types. The first is a short description (15–40 words) of the movie, while the second is a long storyline, which varies from 30–200 words and contains much more details about the different characters and events of the movie. We have also collected the genres of each film from IMDb. We then divide the plots using a 4-act structure. For scene generation, we take the scripts from IMSDb and annotate them with the key elements of a scene.

### 4.1 Plot Generation Dataset

We have created a dataset of 1000 plots consisting of both Bollywood and Hollywood plots, extracted from Wikipedia using the *wikipedia* module in python. The plots collected are around 700 words long on average.

#### 4.1.1 Annotation Guidelines

We annotate the plots by manually dividing them into 4 parts using the 4-act structure described in [appendix A.5](#). We place a single tag at the end of each act: *<one>* (Act 1), *<two-a>* (Act 2 Part A), *<two-b>* (Act 2 Part B) and *<three>* (Act 3) as delimiters. An example for plot annotation is given in the [appendix \(Figure 6\)](#).

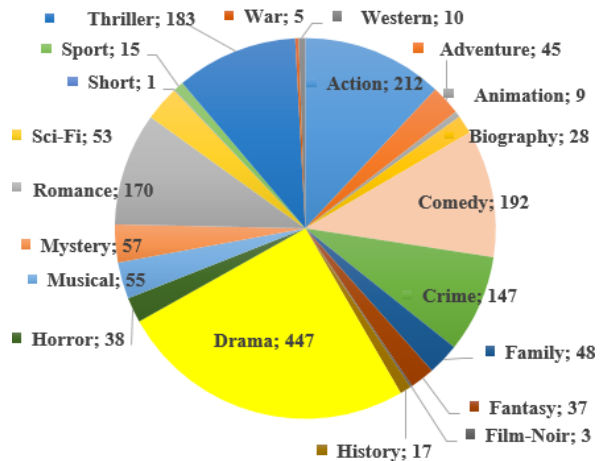


Figure 2: Genre distribution within the plot dataset

### 4.1.2 Movie Genres

To bring some controllability to the plots generated by the model, we have introduced the genres of the movies in the dataset along with the storyline. We concatenate the genres at the beginning of the storyline. Figure 2 shows the distributions of genres in the dataset.

### 4.2 Scene Generation Dataset

Movie scripts are very long. A 2-hour movie corresponds to around 30,000 words. Language models used for creative text generation, like GPT-2 and GPT-3, have token limits of 1024 and 2048, respectively, making it impossible to handle an entire script in one go. Hence, we divided the scripts into scenes and manually created their short descriptions. This allows training the scenes independently instead of relying on any previous scenes.

Movie scripts comprise of multiple elements described in [appendix A.4](#). The different elements increase the difficulty models face in learning to distinguish each element. To overcome this obstacle, we tag four major elements throughout the script — *sluglines*, *action lines*, *dialogues* and *character names*.

#### 4.2.1 Annotation Guidelines

We keep the four major elements present in every script — *sluglines*, *action lines*, *character name* and *dialogues*— and remove any other type of information such as page number, transitions or scene dates. The tagging of the four major elements is done using beginning and ending tags that are wrapped around the elements, as shown below:

- Sluglines: *<bsl>*...*<esl>*

```

<bsl> INT. KENNY'S THAI FOOD DINER - DAY <esl>
<bal>
Kat and Mandella pick apart their pad thai. Mandella is
smoking.
<eal>
<bcn> KAT <ecn>
<bd> So he has this huge raging fit about
Sarah Lawrence and insists that I go to
his male-dominated, puking frat boy,
number one golf team school. I have no
say at all.
<ed>
<bcn> MANDELLA <ecn>
<bd> William would never have gone to a
state school.
<ed>

```

**Figure 3:** The image depicts a portion of a movie scene with the four major elements annotated.

- Action Lines: <bal>...<eal>
- Character Name: <bcn>...<ecn>
- Dialogue: <bd>...<ed>

An example of an annotated scene is seen in Fig 3.

## 5 Experiments and Evaluation

We fine-tune GPT3 with our datasets (refer [appendix A.6](#)).

### 5.1 Plot Generation

We have created 5 models by fine-tuning GPT-3 with our movie plot dataset in the following manner, (i) **original** (without annotation) (**O**): input- short storylines, output- plots without any annotations, (ii) **annotation and short input (AS)**: input- short storylines, output- plots annotated with 4-act structure, (iii) **annotation and long input (AL)**: input- long, more descriptive storylines, output- plots annotated with 4-act structure, (iv) **annotation and short input with genres included (ASG)**: input- short storylines and genre, output- plots annotated with 4-act structure, (v) **annotation and long input with genres included (ALG)**: input- long and more descriptive storylines along with the genre, output- plots annotated with 4-act structure.

For automatic evaluation we use **BLEU** (Papineni et al., 2002), **Perplexity** (Jelinek et al., 1977), **ROUGE** (Lin, 2004). We also use human evaluation in the form of a five-point Likert Scale (Likert, 1932). The rating system has 1-> Strongly Disagree, 2-> Disagree, 3-> Neutral, 4-> Agree, 5-> Strongly Agree. Human-written stories are assumed to have a rating of 5 for each of the following 5 features: (1) **Fluency**: grammatical correctness; (2) **Coherence**: logical ordering of sentences and paragraphs; (3) **Relevance**: Whether the key points

from the prompt have been highlighted in the output; (4) **Likability**: The measure of how much the story is enjoyable; (5) **Creativity**: If the output introduced any new events, character profiles, or relationships.

For plot generation, we generate 50 plots from 50 test prompts. We divide the stories into five groups of 10 and assign three evaluators to each group.

For scene generation, we generate ten scenes from 10 test prompts. We assign five evaluators to rate these ten stories.

## 6 Results and Analysis

We present our observations and evaluations. The nature of our task makes human evaluation take precedence over automatic evaluation (it is for automatic movie script generation, after all!). The qualitative analysis of our generated plots and scenes is based on feedback from 5 professional scriptwriters of our industry partner, the well-known media platform.

### 6.1 Plot Generation

#### 6.1.1 Automatic Evaluation

Table 1 shows auto-evaluation scores for the multiple GPT-3 plot generation models.

#### Storyline:

An alien on Earth loses the only device he can use to communicate with his spaceship. His innocent nature and child-like questions force the country to evaluate the impact of religion on its people.

#### Plot:

An alien lands in a field in Lincoln, Nebraska, on a meteorite, and decides to stay in the area. He soon becomes interested in a nearby town, and in a library he encounters the main character, 8-year-old Owen Garvey. After drawing a spaceship on a library floor with a stick, Owen asks the alien what he is. The alien explains that he is from another planet, and that his people are in danger of being destroyed by pollution. Knowing that Earth is populated by many different religious groups, he asks Owen to be his ambassador to Earth and to help him save his people.  
<one>...

**Figure 4:** The above paragraph is a partial example of a movie plot generated by the model fine-tuned with input as short storyline and output as plot annotated with the 4-act structure.

#### 6.1.2 Human Rating

We conducted human evaluation on Hollywood annotated short input model. The evaluation was done by five groups of 3 people, with each group



Metrics	Models	O	AS	ASG	AL	ALG
<b>Perplexity</b>		2.48	<b>1.84</b>	2.43	2.33	2.63
<b>BLEU-2 (%)</b>		12.95	12.01	12.51	13.08	<b>14.52</b>
<b>BLUE-3 (%)</b>		4.70	4.21	4.55	4.84	<b>5.59</b>
<b>BLUE-4 (%)</b>		2.14	1.92	2.13	2.27	<b>2.59</b>
<b>ROUGE-L (%)</b>		22.67	21.72	23	24.02	<b>24.88</b>
<b>Distinct 3-gram (%)</b>		97.55	97.61	97.39	97.28	<b>98.09</b>
<b>Repetition 3-gram (%)</b>		1.99	2.02	<b>1.72</b>	1.89	1.74

**Table 1:** Scores from common evaluation metrics for 5 Hollywood plot generation models fine-tuned on GPT-3 as O, AS, ASG, AL, ALG (5.1)

having been assigned 10 unique plots. The ratings given for the 5 features are in Figure 5. The average scores for fluency, creativity, likability, coherence and relevance are **3.98**, **3.29**, **2.97**, **2.65** and **2.55**, respectively. Fluency of almost 4 is an indicator of the power of GPT-3 as a language model. Creativity and likeability are respectable at a value of around 3.0. The low BLEU scores support the average creativity score (Table 1). Figure 5 indicates that coherence and relevance still have major room for improvement.

The MAUVE (Pillutla et al., 2021) value measures the gap between neural text and human text. We have separately calculated the MAUVE scores for 20 plots and 50 plots. The weighted average of the MAUVE scores for the two experiments is **0.48** which is reasonably good.

### 6.1.3 Qualitative Observations

Professional scriptwriters from our industry partner have given the following observations:

#### Non-annotated Hollywood Plots

- The build-up is creative and interesting, but the ending becomes incoherent.
- Some characters which are introduced in the beginning are never mentioned again.
- The output is not portraying the key points or the theme mentioned in the input.

#### Annotated Hollywood Plots

- The plots are much more coherent, and the endings are logical.
- There is still hallucination present (a common feature of all models).
- The longer inputs made the plots more attentive to the key points.

#### Annotated Hollywood Plots with Genres included

- Along with the above points, now the plots generated are more tilted towards the genre or genres of the movie the writer wants to create.
- Addition of genre gives some control over the kind of plot generated by the model.

#### Annotated Bollywood plots

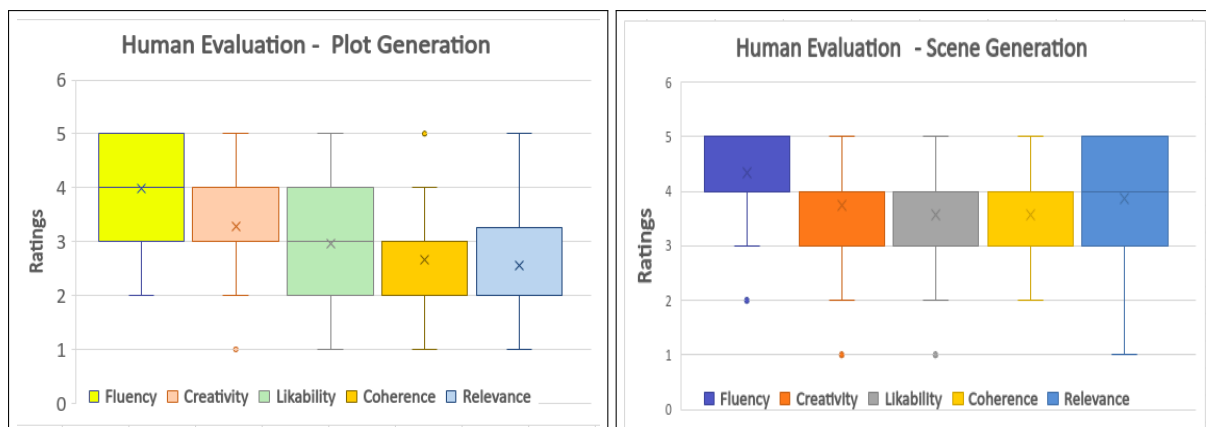
- The outputs show incoherence in the last two paragraphs and repetition of the same characters throughout the plot.
- The flow of the plot is not fast enough, i.e., the plot does not move ahead much.
- Many of the outputs have a 1990s theme around them, where the characters are separated and then find each other later. This is due to a skewed dataset with lesser modern plots.

## 6.2 Scene Generation

We fine-tuned GPT-3 for scene generation with our dataset. We generated ten scenes using the models mentioned in 5.1. Figure 7 in the appendix. shows an example of a completely generated scene.

### 6.2.1 Human Ratings

We conducted a human evaluation on 10 scenes generated by the above model. 5 people evaluated the scenes using the Likert Scale. The ratings for the five features can be seen in Figure 5. The average scores for *fluency*, *creativity*, *likability*, *coherence*, and *relevance* are **4.48**, **3.9**, **3.48**, **3.46** and **3.86**, respectively. All of the values are above the neutral mark and imply that the generated scenes are close to human-written scenes.



**Figure 5:** Boxplot graphs for Human Evaluation of the plot and scene generation models.

## 6.2.2 Qualitative Observations

In this section, we analyze the quality of the scenes generated by the GPT-3 model. This analysis has been done by professional scriptwriters from the previously mentioned media company.

- The model produces a well-structured scene.
- It can create new characters and fabricate dialogues even when they are unimportant.
- The key points from the input can be found in the output.
- There are some lines that are repetitive.
- The output is not completely coherent.

## 7 Conclusion and Future Work

In this paper, we have reported a first-of-its-kind work on automatic plot and script generation from prompts. Automatic evaluation, human rating using the Likert scale, and qualitative observations by professional scriptwriters from our industry partner (a large and well-reputed media platform)- all vindicate the power of our rich dataset and GPT3 in script generation. We hope our work will help television show writers, game show writers, and so on.

There are several future directions: (i) the imbalance in the Bollywood plot dataset needs to be rectified; (ii) there is a lot of variation in Indian script because of multilingualism, which needs addressing; (iii) the most obvious weakness of GPT-3 is not being able to handle factual data and numbers, causing hallucination and preventing the automatic generation of documentaries and biographies. Detection and resolution of hallucination is anyway a growing need for language models.

## 8 Limitations

- In the plot generation dataset, the Wikipedia plots are sometimes not written by professional content writers from the film industry. Therefore these plots may fail to include the main events of the movie.
- In a few cases, the model fails to generate coherent events along with the abrupt introduction of characters in the plots and scenes.
- Although it has been noticed only a few times, the plot or scene generated contains repeated clauses or phrases.
- The model hallucinates and generates factually incorrect things, making it incapable of generating biographies or documentaries.
- The plot or scene may not abide by the theme of the input or genre mentioned along with the prompt.

## References

- Selmer Bringsjord and David Ferrucci. 1999. *Artificial intelligence and literary creativity: Inside the mind of brutus, a storytelling machine*. Psychology Press.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*.

- S. Field. 1979. *Screenplay: The Foundations of Screenwriting*. A Delta book. Dell Publishing Company.
- Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54.
- Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.
- Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *arXiv preprint arXiv:1707.05501*.
- Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and J. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62.
- Mihir Kale and Abhinav Rastogi. 2020. Template guided text generation for task-oriented dialogue. *arXiv preprint arXiv:2004.15006*.
- Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.
- Rensis Likert. 1932. *A technique for the measurement of attitudes / by Rensis Likert*. Archives of psychology ; no. 140. [s.n.], New York.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Rafael Pérez Y Pérez and Mike Sharples. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(2):119–139.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.
- Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.
- Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. **Target-guided open-domain conversation**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.
- Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*.
- Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. **Proactive human-machine conversation with explicit conversation goal**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.
- Rui Yan. 2016. i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In *IJCAI*, volume 2238, page 2244.
- Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.
- Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun Nie, and Jin Zhou. 2020. **ScriptWriter: Narrative-guided script generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8647–8657, Online. Association for Computational Linguistics.

## A Appendix

### A.1 Ethics Consideration

We have taken all the scripts from IMDB and IMSDb databases. The website has a disclaimer regarding using its scripts for research, which can be found at this link <https://imsdb.com/disclaimer.html>. We have used the scripts fairly and without copyright violation.

### A.2 Annotator Profiles

We required the help of external annotators in two cases: (i) Manually Annotating the Scripts and (ii) Creating scenes and their descriptions from the scripts. For the first task, we took the help of 10 annotators. Their ages ranged from 21–28, and all were Asian. They were given detailed guidelines

with examples for annotating. There were also periodic sessions to confirm their understanding and solve their doubts and mistakes. For the second task, we took the help of two annotators. Both of them are Asian females aged between 21-23. Both of them were given detailed guidelines for the scene-writing task. A few data points were picked randomly and checked to find out and correct conceptual mistakes. The annotators had bachelors and masters degree in STEM and Arts.

### A.3 Evaluation Metrics

The evaluation metrics are described below:

- **Perplexity (PPL):** Perplexity is one of the most common metrics for evaluating language models. They are computed as exponential of entropy. The smaller the value of the PPL, the greater the fluency of the generated text.
- **BLEU: BiLingual Evaluation Understudy** is a common metric in many NLP tasks, especially in the field of Machine Translation. It measures the overlap between the generated output and gold standard data. Although this metric does not consider the model's creativity, we can deduce the difference between the candidate text and the reference text using BLEU. The higher the BLEU measure, the better it is.
- **ROUGE: Recall-Oriented Understudy for Gisting Evaluation** is typically used for evaluating automatic summarization. In our case, it measures the longest overlapping sequence between the generated and original plots. The higher the ROUGE measure, the better it is.
- **N-grams:** We measure the redundancy and diversity of the movie plots by computing the repetition and distinction n-gram scores.

### A.4 Screenplay Structure

A movie script or a screenplay has a different format than a story. A script is a group of scenes. Each of these scenes consists of a few major components, which are discussed below:

**Scene Headings/Sluglines-** This component describes the when and where of the scene. It can be thought of as the first shot that a camera takes of a new scene. For example, INT. - RESTAURANT - NIGHT indicates that the scene starts inside a restaurant at night. Sluglines are normally written in capital letters and are left-aligned.

**Character Names-** they are mentioned every time a character is going to utter a dialogue. The name of each character is mentioned in uppercase and is centre aligned.

**Dialogues-** dialogues are the lines that the characters say. They appear right after the character name in a script and are centrally aligned.

**Action Lines-** action lines describe almost everything about a scene. They can be described as the narration of each script. Action lines can be present after either dialogues or sluglines and are left-aligned.

**Transitions-** a transition marks the change from one scene to the next. They also depict how a scene is ended. For example, DISSOLVE, FADE, and CUT are different keywords used to indicate a transition. They are usually in upper case and are right-aligned.

Figure 8 shows an example of the screenplay elements.

### A.5 Story Templates

Over time various templates have been developed that help to create stories. One of the most famous templates is the 3-act structure (Field, 1979). This structure divides a story into a *setup, confrontation, and resolution*. In this work, we have used the 4-act structure which we now describe in detail.

**Act 1-** This is the opening/introduction act. It describes the protagonist's character and briefly introduces the movie's theme. The act ends with the start of a new journey for the protagonist.

**Act 2A-** Due to the vast span of Act 2, it can be divided into two acts. This act usually contains the start of a love story. It also entertains the audience as the protagonist tries to adapt to their new journey. The act ends as the movie's midpoint, one of the film's critical moments, with either a very positive or negative scene.

**Act 2B-** This act usually contains the protagonist's downfall. The villain or antagonist starts to gain an advantage, and the protagonist loses something or someone significant. The act ends with the protagonist realizing their new mission after reaching rock bottom.

**Act 3—** The protagonist has realized the change required in them and sets out to defeat the antagonist in a thrilling finale. The movie then ends by displaying a welcome change in the protagonist that was lacking in the beginning.



In 1988 New York City, Roberta Guaspari, a recently divorced violinist, lives with her two sons, Alexi and Nicholas Tzavaras, and her mother, Assunta Guaspari. With Assunta's encouragement, Guaspari attempts to rebuild her life and is recommended to the head teacher of East Harlem's Central Park East School. <one>

Despite having little experience in actual music teaching, she accepts a substitute violin teaching position at Central Park East. With a combination of her toughness and determination, she inspires a group of children, and their initially skeptical parents. The program slowly develops and attracts publicity, eventually expanding to Central Park East II and River East Schools. <two-a>

Ten years later, the Central Park East, Central Park East II and River East School string programs work with the New York City Board of Education to help eliminate funding for the programs, which leads to Guaspari's early dismissal. Determined to fight the budget cuts, she enlists the support of former pupils, parents and teachers and plans a benefit concert, Fiddlefest, to raise money so that the program can continue. But with a few weeks to go and all participants furiously rehearsing, they lose the venue. <two-b>

However, Arnold Steinhardt, the husband of a publicist friend, is a violinist in the Guarneri Quartet, and he enlists the support of other well-known musicians, including Isaac Stern and Itzhak Perlman. They arrange for the concert to be mounted at Carnegie Hall. On the day of Fiddlefest, Guaspari and her students perform with Perlman, Steinhardt, Stern, Mark O'Connor, Michael Tree, Charles Veal Jr., Karen Briggs, Sandra Park, Diane Monroe, and Joshua Bell, increasing donations and making the event a massive success. In the epilogue, descriptions show Guaspari and the Opus 118 program's activities after the events in 1991. <three>

**Figure 6:** Example of manual annotation of the plot of the movie *Music of the Heart* using the 4-act structure

## A.6 Fine-Tuning GPT-3

GPT-3 was deemed publicly available last year by OpenAI (Brown et al., 2020). Its best model has 175B parameters, which is much more than GPT-2's 2.9B parameters. We have fine-tuned multiple plot generation models with GPT-3 along with a scene generation model. The multiple combinations of plot generation models are short or long prompts and with or without genres. The GPT-3 model and hyperparameters remain the same for all the above combinations. We have fine-tuned the GPT-3 Curie model for four epochs. For generating text, GPT-3 offers various hyperparameters to tune and get closer to our desired results. For testing, we set other hyperparameters as follows: the temperature as 0.7, top-p as 1, frequency penalty as 0.1, presence penalty as 0.1, and max tokens as 900.

**Scene Description:**

Jack goes to an old town and finds himself next to an old and rusty haunted mansion. He goes inside and tries to describe the mansion.

**Scene Generated:**

EXT. FOREST - TEARDROP NIGHT

Jack goes to an old town and finds himself next to an old and rusty haunted mansion. He goes inside and tries to describe the mansion.

INT. THE MANSION - TEARDROP NIGHT

Jack is in a huge old house. He is still trying to describe the house.

JACK

There are dark, musty rooms, with ghostly furniture and dusty old antiques. It's filled with old, creaking things, I don't know if it's set on the North, South or East side.

WENDI

It's on the west side.

JACK

Oh, good. Oh, great. Well I tried, I tried my best Wendi.

Wendi pats Jack's back.

WENDI

You did good. Now you're a regular King of Darkness.

Wendi takes off her glasses and rubs Jack's back some more. She puts her glasses back on and looks at Jack.

WENDI

Listen, I'm going to have to ask you to leave now. I have some business to handle, okay?

Jack tries to leave but Wendi stops him again. She opens a door and we see a bunch of candles and wax drippings on the floor. Wendi steps over the wax and leads Jack through the secret passageway. Jack looks at her. She is beautiful in the light of the flickering candles. They walk a bit more, then Wendi turns off the hall light, closes the door and locks it. They hear a noise outside the door and Wendi goes to check it. She opens it again and they are back out in the hall. They walk a little more down the hall and stop at another door. Wendi opens it and they are back outside in the dark night woods.

WENDI

Keep going straight and then turn right at the first street you see.

Wendi shuts the door and locks it again. Jack is alone in a big old dark scary house. He walks down the dark hallway toward the secret passageway. He looks at some old paintings of old women on the wall, with very bad haircuts and clothes that no one would wear today. He walks through some strange hallways, past some more weird paintings, past some scary paintings of old people with skeletons in their mouths and heads on sticks...

**Figure 7:** An example of a complete scene generated given a short input.

INT. COFFEE SHOP – MORNING

Scene heading

A normal Denny's, Spires-like coffee shop in Los Angeles. It's about 9:00 in the morning. While the place isn't jammed, there's a healthy number of people drinking coffee, munching on bacon and eating eggs.

Action

Two of these people are a YOUNG MAN and a YOUNG WOMAN. The Young Man has a slight working-class English accent and, like his fellow countryman, smokes cigarettes like they're going out of style.

It is impossible to tell where the Young Woman is from or how old she is; everything she does contradicts something she did. The boy and girl sit in a booth. Their dialogue is to be said in a rapid pace "HIS GIRL FRIDAY" fashion.

YOUNG MAN

Character

No, forget it, it's too risky. I'm through doin' that shit.

Dialogue

YOUNG WOMAN

You always say that, the same thing every time: never again, I'm through, too dangerous.

Figure 8: The elements of a screenplay