[1]Vaishali Ganganwar*
[2]Divyanshu Gupta
[3]Vinay Kumar Singh
[4]Jayanth
[5]Abhishek Thakur

# Experiments with Storytelling-Improving Task-Specific Performance in GPT-2


***Abstract:*** - Creating engaging stories is an inherently human art, and the possibility of a computational model being able to mimic this task has long been a topic of discussion as it requires the unique intersection of the fields of artificial intelligence, psychology and literature. Computational generation of stories helps reduce efforts, find inspiration, and tailor stories to the user's education or entertainment needs. This research focuses on performing several independent experiments with the base GPT-2 model, in an effort to improve its ability to generate stories from a prompt. The three major experiments conducted are double fine-tuning, addition of commonsense knowledgebase, and sequential transformers. Additionally, the research provides insights on the effects of double fine-tuning and addition of knowledge bases to a large language model such as GPT-2. All experimental setups are evaluated on metrics such as perplexity, BLEU and ROGUE, in order to give an idea of their performance in quantitative measures. The best performance across genres is observed on the double fine-tuned model with the knowledgebase added. The study concludes with insights on possible improvements to the GPT-2 model that can help improve story generation efforts, as well as suggestions on improvements of the evaluation framework available for evaluation of creative work performed by AI.

***Keywords:*** Story generation, GPT-2, Pretrained Transformer, longform-text

## I. INTRODUCTION

A story can be defined as a medium, written or spoken, for sharing information, experiences, emotions, values, and much more. A story can be used to achieve many objectives, like entertainment, education, delivery of meaningful experiences, etc., and is filled with engaging characters and an immersive plot, which grabs the reader's attention. Before considering the intriguing domain of Automatic Story Generation, it is essential to start by defining the essential components of a story. A story constitutes of various elements and considerations, like thematic narrative, character development, plot progression, linguistic expression, literary devices, and most importantly, artistic influence. All of these elements are painstakingly combined by a writer, to create a cohesive, engaging and complete story.

Coming to the domain of automatic story generation, as it has evolved over decades, three overarching techniques have primarily been researched and developed upon: Structural Models, Planning-Based Models, and Machine Learning (ML) Models. Structural Models work using the definition of a story, and the independent manipulation of its structure. These are further categorized into Graph-Based and Grammar-Based Models. Planning-Based Models were introduced with the advent of Artificial Intelligence (AI) technologies. These models include Goal-directed Models in which the means are used to define the end. This framework works by establishing an objective and utilizes reinforcement learning by evaluating each generated story and rewarding exemplary narratives. Goal-based stories prioritize character needs over grammatical considerations. Another type of Planning-Based models in the Heuristic Search approach. A genetic algorithm is used to generate stories, by dividing them into timestamps represented by story components. Each component is encoded into a chromosome, with the genes corresponding to story elements, such as agents and objects. Lastly, a significant rise in the use of large language models has been observed in all sorts of long-text generation tasks, thanks to a boom in the AI market. The majority of these systems use Recurrent Neural Networks (RNNs). With the addition LSTM (Long Short-Term Memory) systems, RNNs have shown the ability to statistically learn and generate long form text.

[1] *Corresponding author: Army Institute of Technology, Pune., India. vganganwar@aitpune.edu.in

[2] Army Institute of Technology, Pune., India.

[3] Army Institute of Technology, Pune., India.

[4] Army Institute of Technology, Pune., India.

[5] Army Institute of Technology, Pune., India.

This research is highly motivated by the ongoing boom in research and development of highly specialized AI-powered systems that do exceptionally well in hyper-specific tasks, such as programming, music, medical sciences, law consultation, etc. Story generation was chosen as the domain for implementing such a system, as it is an incredibly intriguing and complex human art form, that is tough to fully replicate as it includes elements such as characters, motivations, emotions, evolving relationships, etc. Additionally, stories are a curious case study when considering long-text generation systems, as the text has to hit a certain word-limit, but still keep relevance to the plot. Finally, researching this domain using GPT-2 allows us to deeply understand the transformer architecture and how it may be improved for future implementations, while giving a clear understanding of the limitations of using a model that is much smaller compared to the proprietary implementations available in the commercial domain, namely GPT-3.5, GPT-4, Gemini, etc.

This research comes after an extensive survey of the literature on the development of story-generation systems using resources available in the open-source domain. It has been identified by the authors that the GPT-2 model is the best large-language model that works on the pre-trained transformer architecture, that can be used as a base for extrapolating improvements to the transformer architecture. The research comprises of four different experimental setups to evaluate the performance of GPT-2 for story generation and identify the ideal methodologies to improve said performance. The datasets used in the research are a combination of the state-of-the-art datasets and an assortment of web scraped stories for each genre.

The key contributions of our paper include an in-depth analysis of GPT-2 performance across four different story generation setups, as well as tabulation of automated scores of one-shot story generation by GPT when used with each of those modifications. This will be helpful for anyone looking for insights on GPT-2 and its ability to write stories. The human evaluation provided acts as an added source of qualitative information about the capabilities of the model as well as the nuances of each dataset.

## II. Literature Survey

Research in this area of natural language generation has been spread across various methodologies, datasets, and the fundamental setup of the model in the number and types of inputs and outputs it is supposed to provide. To organize our study of the available literature on this subject, we group together models that take similar inputs, have similar levels of user interaction, and provide similar outputs. We further group the research on the basis of the base methodology they use.

Fan et al., 2018 [5] employed a comprehensive methodological approach to generate coherent stories from writing prompts. They gathered a substantial dataset of 300,000 human-written stories and corresponding prompts from an online forum and used a novel hierarchical story generation method that initially created a premise and then converted it into a complete passage. Comparing their approach to a KNN model using TF-IDF vectors demonstrates the superiority of this method. The top-k random sampling scheme for generating stories outperformed traditional beam search. Results revealed that the hierarchical approach was significantly preferred by human judges and exhibited substantial improvements over non-hierarchical models in both automated and human evaluations. There are however limitations such as dataset biases, resource requirements, generalizability to different domains, ethical considerations, and even considering the comparison that we did among other extremely sophisticated and efficient models. Building on this, Chen et al., 2021 [2] introduced a neural model for ending of story selection model that effectively integrates narrative sequence, sentiment evolution and commonsense knowledge, and surpasses existing best approaches on the ROCStory Cloze Task, achieving significant improvements through the incorporation of additional commonsense knowledge. It also discusses neural SeqMANN models that achieve high accuracy by combining various features, including embeddings, character features, sentiment polarity, part of speech tagging features, negation information, and external semantic sequence knowledge. Pre-training word embeddings on a large external corpus further enhance model performance. Discussion regarding limitations or shortcomings in the pre-training and fine-tuning process of the language model used however are pending and of much concern. Ammanabrolu et al., 2019 [1] also goes on to propose an ensemble-based model for natural language generation guided by events, which outperforms the existing seq2seq model. The paper mentions the use of a finite state machine constrained beam search as the final model for sentence generation. This method involves training a seq2seq model on pairs of events and generalized sentences and using Monte Carlo beam search as an alternative search strategy within the decoder network.

Yao et al. 2019 [26] and Rashkin et al., 2020 [17] focuses on generating an intermediate outline before proceeding to the full-fledged task of story generation. The first group used a plan and write framework consisting of two main strategies: the dynamic schema and the static schema. Going over dynamic scheme, what it does is, it interweaves story planning and its surface realization in text. Contrary to dynamic schema, static schema plans out the whole story line before generating stories. The other group proposes Plot-Machines, a narrative model based on neural network, for outline-conditioned story generation. This model tracks the dynamic plot states, so that it can go on to transform the outlines of the story into some coherent story. The model contains within itself, high level discourse structure. This high level of discourse structure helps the model learn different styles of writing that correspond to different parts of narrative. Experiments compared the performance of Plot-Machines with large scale language models such as GPT-2 and Grover and found that these language models, despite their impressive generation performance, were not good at generating coherent narratives for any given outline.

A general trend of automated story generation is that the models focus either on sentimental consistency [9] and logic, or on commonsense logic to maintain a consistent narrative. Both approaches tend to forgo the other aspect of storytelling, resulting in an overall inconsistent result. Guan et al., 2020 [11] also propose a knowledge-enhanced pre-training model for commonsense story generation, which as the name suggests, utilizes commonsense knowledge from external knowledge bases to generate reasonable stories. In contrast to this, Mo et al., 2021 [13] propose a Gated Mechanism based Transformer Network (GMTF) for story-ending generation. The GMTF model incorporates the sentimental trend into the story-ending generation process to make it more consistent with the sentiments present in the text [8]. The sentimental trend is obtained using a sentiment analysis tool called VADER. The use of sentiment analysis [6] and contextual information analysis [7] is that it goes on to help us capture key clues, once we enter them into our transformer network. This results in stories with better sentimental consistency than other models. In a more novel pursuit, Cheong et al., 2015 [4] showed the development of a system that produces suspenseful narratives by manipulating the reader's suspense level through the story structure. Using a plan-based approach for narrative comprehension, the final content of the story is generated. It tries to keep the suspense level for the reader through its generation. The architecture of the system is designed as a three-stage pipelined architecture, consisting of a fabula generator, sjuzhet generator, and discourse.

A lot of research has also been done with regard to providing more granular control of the story-generation process to the use, through various interactive tools. Goldfarb-Tarrant et al., 2019 [10] goes on to present a generation system that is based on neural network that interacts with humans and then generates story based on those inputs. Varied levels of human interaction were shown in the system, allowing for understanding the most productive stage of story writing collaboration. The authors compare different varieties of interaction in story-writing, story-planning, and diversity controls under time constraints. J.Chen et al., 2021 [3] Control and edit transformer technique was introduced by the authors for story plot generation. Controlled imitation was key feature for editing distance from dynamic programming, so that policies like deleting, supporting, inserting can be supported. A weight reward with preprocessed corpus statistics and measure continuous reward for the controlled goal were also supported. While both efforts provide some level of control on story generation, fine-grained control is offered by Wang et al., 2022 [22] with their model called CHAE for story generation, allowing the generation of customized stories, and emotions arbitrarily assigned and placed as conditionals at every point in story generation. The model however is tested only on the ROCStories dataset, and further investigation and testing is needed to confirm the usability of the model. Other notable approaches include Xie et al., 2022 [24] and their attempts at using contrastive learning as the method to train the model in order to obtain stories with more consistency and rationality using two major components: multi-aspect sampling and story-specific contrastive learning. Wilmot et al., 2021 [23] also introduced a variational autoencoder which considers the Temporal Difference between different events in a story, known as TD-VAE. The TD-VAE model is used to improve plot development and long-term coherence in story generation. The performance of the TD-VAE model is evaluated using automatic cloze and swapping evaluations, which demonstrate strong performance. The Recursive Reprompting and Revision framework (Re3) was proposed by Yang et al., 2022 [25] to generate longer stories by prompting a general-purpose language model to construct a structured overarching plan and generating story passages by injecting contextual information from both the plan and current story state into a language model prompt. This increased prompt specific relevance and long-term coherence by substantial amounts. Tang et al., 2022 [20] present the MVP model - Multi-task superVised Pre-training, for natural language generation that works by pre-training task-specific prompts by prepending trainable vectors to multi-head attention modules at each layer, and also provide a new dataset named the MVPCorpus to

assist in research in this field. In the realm of language models, Brown et al., 2020[21] explore the capabilities of GPT-3, a language model with 175 billion parameters, in various NLP tasks without fine-tuning. GPT-3 demonstrates strong performance across multiple tasks, highlighting its potential as a few-shot learner for narrative generation and other NLP applications. Representing the latest advancement in narrative generation Saravanan et al., 2022[19] harnessed the power of GPT-3 for autonomous content generation and transformation. By leveraging GPT-3's language prediction capabilities, this system explores the potential of generating and transforming content without manual human intervention, paving the way for further advancements in automated storytelling and content creation.

Four major datasets are available for the task of training or fine tuning a model for the task of story generation: ROCStories, WritingPrompts, Hippocorpus and Wikiplots. Understanding narratives and extracting implicit knowledge from text is a fundamental challenge in natural language processing. The ROCStories dataset was introduced by Mostafazadeh et al., 2016 [14] to address this challenge by providing a rich collection of commonsense-based short stories. The ROCStories corpus comprises 100,000 stories, each of them five sentences in length. These stories were logically constructed around mundane topics by Amazon Mechanical Turk workers. Each story captures a variety of essential components, most notably commonsense causality and time-based relations between everyday events. the writers also contributed an additional set of 3,742 stories that make up the Story Cloze Test, each consisting of a body four sentences long that gives the required context, and two candidate endings. The published ROCStories dataset consists of three main components: the Training Set: Comprising 98,162 stories, excluding candidate wrong endings; the Evaluation Set: Structured similarly to the training set, with a size of 1,871; the Test Set: Also following the same structure of one body and two candidate endings, with a size of 1,871. The WritingPrompts dataset introduced by Fan et al., 2018 [5] is a huge collection of human-written stories each associated with their respective writing prompts scraped from the WritingPrompts subreddit, a community on the popular online forum, Reddit. Variety of genre, theme, detail and lengths have been observed in datasets of Writingprompt. The dataset is a valuable resource for researchers working on story generation models and is split in the following manner: 272,600 training stories, 15,138 stories for testing, and 15,620 stories for validation. The average length of the prompts is 28.4 tokens, and the average length of the stories is 734.5 tokens.

The Wikiplots dataset is extracted from various Wikipedia pages of stories, movies and TV shows. The dataset is a collection of 112,936 synopses of movies, books, and other works of fiction. Each synopsis is a brief summary of the plot of the work and is typically only a few sentences long. The Hippocorpus dataset introduced by Sap et al. [18]is a series of diary-like short stories of salient life-events. The dataset has 6,854 total stories, collected from Amazon Mechanical Turk. In the first stage, recalled stories were collected, where workers wrote a 15-25 sentence story about a memorable event that they had experienced in the last 6 months. There are 2,779 such stories in the dataset. In the second stage, a new set of workers were chosen to write a story based on a randomly assigned summary from the previous stage. There are 2,756 such stories. The third and final stage comprised of the workers from the first stage retelling the stories they told earlier using the summary they provided then, to test the impact of temporal distance and the act of retelling on storytelling. There are 1,319 such stories.

## III. METHODOLOGIES

The experiments in this research were done by using the base GPT-2 model in various configurations, with or without the addition of a commonsense knowledge base. These experimental configurations and modifications are described in this section.

### *A. GPT-2*

GPT-2 is the second version of the Generative Pretrained Transformer which is a natural language processing model by OpenAI [16]. It is based on the transformer which is considered as the standard for natural processing models nowadays. Transformers are the type of neural networks best at handling sequential data which makes them suitable for different tasks like text generation , machine translations and language modeling. This is model which is pretrained on large text data taken from internet. This model is trained and evaluated against the data called WebText. A WebText dataset comprises a diverse collection of text data sourced from the internet, covering an array of topics, styles, and formats. It includes content extracted from web pages, social media posts, forums, reviews, news articles, blogs, personal websites, and more. The dataset captures the richness of online communication, incorporating formal and informal language, user generated discussions, opinions, and sentiments.

This amalgamation of textual data provides a valuable resource for natural language processing tasks, such as sentiment analysis, language modeling, and text classification. Researchers often create or compile WebText datasets to analyze and understand the dynamics of online content, but it's essential to approach the collection and usage of such data ethically and in compliance with legal standards.

Thus, the model trains in a self-supervised fashion, using an automated process to label given texts without any human intervention. The model incorporates an internal masking mechanism, ensuring that predictions for token i solely rely on input information from positions 1 to i, preventing any influence from future tokens on the output. This helps the model learn grammar, context, and world knowledge, while fine-tuning adapts it to particular applications like text completion, translation, summarization, and question answering.

The scale, size and variety of the GPT-2 model implementations is one of its most impressive features. The model was released in four sizes based on size like basic, medium, large and XL. The largest model has 1.5 billion parameters to train on and this makes it one among the largest language model. The architecture of GPT-2 transformer is shown in Figure 1. The expansive size of GPT-2 enables it to comprehensively grasp and replicate intricate linguistic patterns and context, facilitating the generation of coherent and contextually relevant text across diverse domains, ranging from creative writing to technical content. Additionally, the model exhibits robust transfer learning capabilities, allowing effective fine-tuning for specific tasks with relatively small amounts of task-specific data. This adaptability makes GPT-2 suitable for a wide array of applications, including text completion, generation, summarization, translation, question-answering, and more. Notably, these capabilities are achieved without necessitating significant architectural changes, rendering GPT-2 an ideal candidate for inclusion in this comparative study.

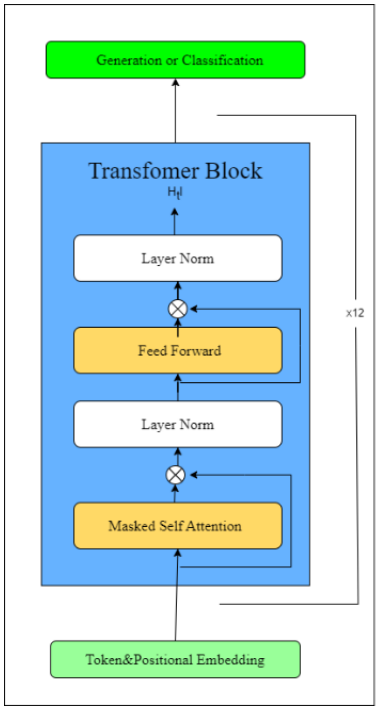


Fig. 1: GPT-2 Transformer Architecture

## IV. EXPERIMENTS

### *A. Double Fine-Tuning*

The GPT-2 model supports fine-tuning using datasets or reinforcement learning as methods of fine-tuning the model for a specific task. Fine-tuning enables already fully trained models to increase their accuracy in a specific task by slightly modifying the internal parameters to better fit a given data distribution. Here, the fine tuning is done using state of the art story generation datasets. This already improves the ability of GPT-2 to generate stories. Once this first stage of fine-tuning is done, the model is again fine-tuned on genre-specific datasets. The process of fine-

tuning and double fine-tuning is described in Figure 2 and 3 respectively. For the purpose of this research, web scraped datasets of Horror, Science-Fiction, and Comedy (Humor) were used, but the results observed can be extended to other genres as well, depending on the availability of data. For the purpose of our research, we chose the model trained on the ROCStories dataset as the base model for the second fine-tuning stage.


Fig. 2: Fine-tuning the GPT2 model

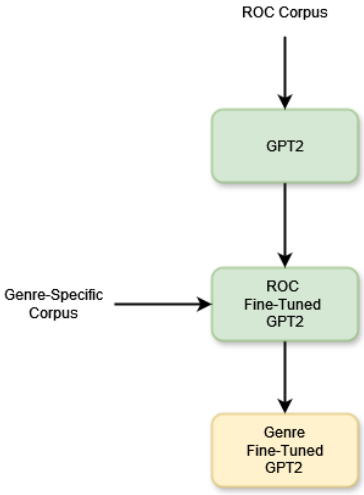


Fig. 3: Double fine-tuning the GPT2 Model

*B. Knowledge-Base Addition*

The addition of knowledge-base is a common technique used in artificial intelligence modelling to improve the model's domain specific knowledge and performance. Upon analysis of stories from the single fine-tuned or double fine-tuned models, it is observable that the GPT-2 model can be improved for this task by improving the commonsense knowledge, as the stories are lacking in general world knowledge. For this purpose, two knowledge bases of commonsense domain were added to the double fine-tuned models, and the performance was evaluated across the same metrics. Figure 4 shows the process of addition of knowledge-base while fine-tuning of GPT2 model. The knowledge bases used were ConceptNet and Atomic, as discussed in the later sections.

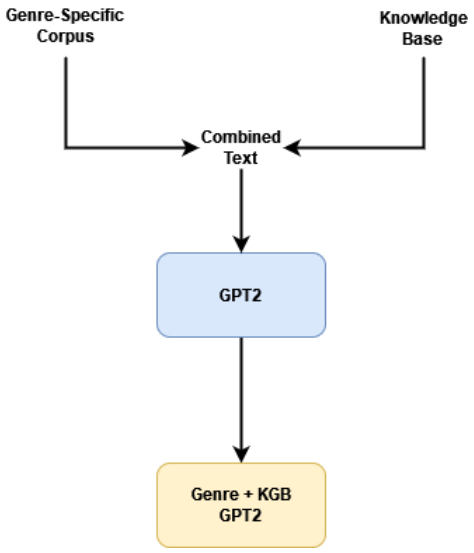


Fig. 4: Addition of knowledge-base to GPT2 Model

### *C. Sequential GPT-2*

A novel approach attempted in this research, ”Sequential GPT-2” is an implementation that uses two GPT-2 models in a sequential arrangement such that the initial prompt for story generation is given to the first model, and the output is truncated to the first 100 words. This output is then used as the prompt to the second GPT-2 model in the sequence, and the final output is obtained. Figure 5 shows the approach. This approach was developed as an attempt to solve a glaring problem with the GPT-2 model, which was its ability to stick to the context of the prompt. a single GPT-2 model could only, on average, give coherent and relevant outputs of story generation for the first 100 words of text. Beyond that, the model would start to lose semantic sense, and absolutely abandon the plotline. The proposed sequential implementation allows the model to reiterate over the already generated text and generate the extended text with more context. This sequential approach was tried with both the versions of the model, the model with knowledgebase added, and the one without.

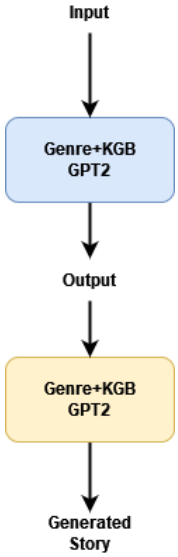


Fig. 5: Sequential GP2 model

## V. DATASETS

The research uses four genre-neutral SOTA datasets to fine tune the model once, and then genre specific datasets for double fine-tuning. The datasets are:

### *A. Sci-Fi dataset*

The Sci-Fi dataset is a dataset of science-fiction stories scraped from the web. The total dataset has 1,109,085 words, in 16,517 lines. It is used to double finetune the model in order to generate genre-specific stories. The dataset is freely available for use on the HuggingFace library. It allows the user to access the stories in the form of a text file, which can be collated to form a bag of words, as we have done in our implementation.

### *B. Scary dataset*

This dataset is a collection of short horror stories, used to fine-tune the model for horror genre. It has 58,000 words, distributed over 950 lines. Also available on HuggingFace and on GitHub, this dataset is the best resource available in the open source. A web scraped dataset can also be made to improve the size and robustness of this dataset.

### *C. Humour Dataset*

A dataset used to train models for the comedy genre, this dataset features 200,000 jokes that can be used to fine-tune the model for a comedic genre. This is also a web-scraped dataset, available on Kaggle and HuggingFace libraries.

## VI. KNOWLEDGE BASES

For adding commonsense knowledge regarding real-world facts as information for the model, two knowledgebases have been used:

### *A. ConceptNet*

ConceptNet is one of the largest free resources available to impart knowledge to models about entities and their relationships. It was developed and launched by the MIT Media Lab in 1999 as part of a crowdsourcing project named Open Mind Common Sense. The knowledgebase is available in the form of a semantic net, that has nodes that represent entities in real life, and edges that represent the relationships between them. It can be effectively used to create word embeddings and can even be used in multilingual implementations.

### *B. Atomic*

The Atomic Knowledge Base is a repository of structured data designed to enhance performance of AI models, particularly in natural language understanding and reasoning tasks. Developed at OpenAI, the Atomic Knowledge Base uniquely encoded knowledge in the form of atomic facts, each representing a single, discrete piece of information. This approach allows for efficient storage, retrieval, and inference of knowledge, enabling AI systems to make nuanced and contextually appropriate decisions. By leveraging this comprehensive repository of atomic facts, models can effectively increase understanding of fact-based commonsense knowledge.

## VII. EVALUATION METRICS

In our experiments, we have incorporated perplexity, BLEU [15] and ROUGE [12] as the evaluation metrics for the model. BLEU provides a quantitative measure of the overlap between the model-generated text and reference texts, offering insights into precision. ROUGE evaluates content overlap by considering n-gram matches and recall, contributing to a holistic understanding of the generated text's semantic alignment with references. Additionally, we employ perplexity as a metric to measure the model's predictive uncertainty, giving us valuable insights into the fluency and coherence of the generated text. It measures how well a model assigns probabilities to sequences of words. These all metrics are evaluated for different models of gpt-2 proposed in the research done here based on different genres to confirm the trend of change in metrics across these datasets.

## VIII. RESULTS

The genre-based datasets namely Sci-Fi stories, Scary stories and humour stories are used to confirm the results for different metrics. For each of these datasets, results are obtained from four proposed models for metrics like Perplexity, BLEU, ROGUE-1, ROGUE-2 and ROGUE-L which are shown in Table 1. Firstly, the simple dataset based fine-tuned model is evaluated which is used as the base for all the other models for a particular dataset(Table 1, rows-'SciFi', 'Scary', 'Humour') . Next the sequential fine-tuned model of the same dataset is evaluated which is used to maintain the context of the prompt given during story generation and the notable changes can be seen as the perplexity is showing a drop in value which is good sign for the model and also the BLEU and ROUGE metrics are improved showing model is performing better than the base model (Table 1, rows-'Seq SciFi', 'Seq Scary', 'Seq Humour'). For improving the commonsense aspects of a story, the next two models are proposed where we integrate the knowledge base(KGB) with models. For KGB models first we integrated the KGB with base model which is dataset fine-tuned and the changes observed are increase in the perplexity of the model as with the integration of KGB the probability for next word increases as more data is provided to the model to predict from but the change is manageable as the model performs better for the BLEU and ROUGE metrics which proves the improved quality of the stories from the model and the change is significant to prove the integration of KGB overall adds to the performance of the model (Table 1, rows-'SciFi+KGB', 'Scary+KGB', 'Humour+KGB'). Lastly the sequential KGB model is evaluated which deals with shortcoming of perplexity for KGB model as it can help in improving perplexity while it further improves the BLEU and ROUGE metrics as it maintains the context of the prompt better than the normal model (Table 1, rows-'Seq SciFi+KGB', 'Seq Scary+KGB', 'Seq Humour+KGB').

| Model | Perplexity | BLEU | ROGUE-1 | ROGUE-2 | ROGUE-L |
|---|---|---|---|---|---|
| **SciFi** | 46.01 | 10.32 | 26.02 | 10.34 | 23.28 |
| **Seq SciFI** | 44.02 | 10.53 | 21.73 | 12.71 | 21.73 |
| **SiFi+KGB** | 54.66 | 11.38 | 35.46 | 11.56 | 31.20 |
| **Seq SciFi+KGB** | 53.65 | 11.56 | 30.55 | 11.11 | 26.38 |
| **Scary** | 20.50 | 12.93 | 26.31 | 14.28 | 26.31 |
| **Seq Scary** | 20.50 | 17.02 | 37.38 | 17.88 | 37.38 |
| **Scary + KGB** | 31.79 | 00.71 | 26.31 | 14.28 | 09.09 |
| **Seq SciFI+KGB** | 31.79 | 00.67 | 08.10 | 04.59 | 08.10 |
| **Humour** | 30.50 | 12.19 | 18.46 | 11.92 | 18.46 |
| **Seq Humour** | 30.50 | 12.25 | 23.14 | 12.41 | 21.48 |
| **Humour+ KGB** | 30.40 | 12.77 | 24.19 | 13.07 | 24.19 |
| **Seq Humour+ KGB** | 30.40 | 12.91 | 23.62 | 12.16 | 22.04 |

Table 1: Results

The graphs (Fig 6,7,8) give a glimpse of all the evaluation metrics for all the models pertaining to one specific genre. This assists us in determining which specific implementation leads to an increase in a particular score. For example, it is clearly visible in the scores for the Sci-Fi models, the sequential model has the best perplexity score while the ROGUE scores are best for the model with the knowledge-base added. On the other hand, the graphs (Fig 9,10,11,12,13) represent the scores that every model within the experiment got for each evaluation metric. These graphs help us better judge the best and worst performing models for each metric, as well as any visible trends. For example, the plot of ROGUE 2 vs the models shows that the basic Sci-Fi model gives us the best score, indicating that some aspect of the models' ability to generate text is actually negatively impacted in all of our experiments.

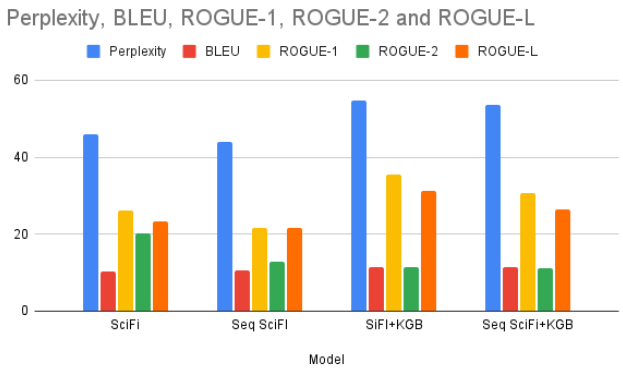


Fig. 6: All Scores for Sci-Fi genre models

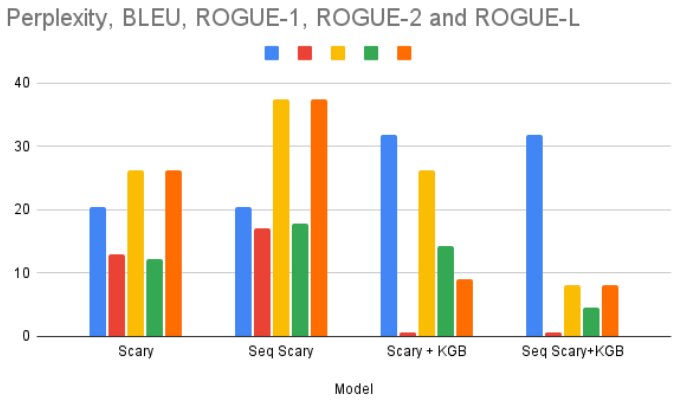


Fig. 7: All scores for Horror Genre models

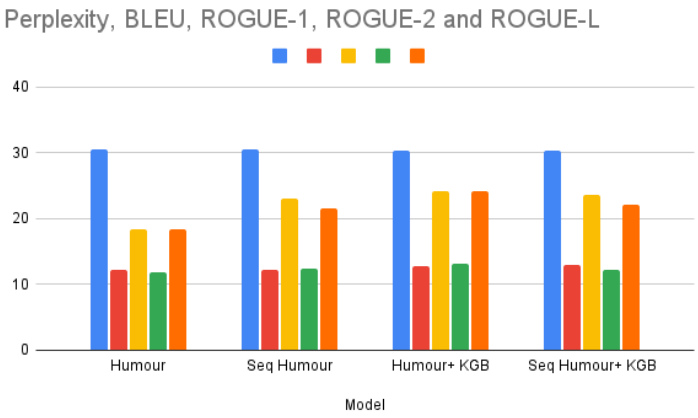


Fig. 8: All scores for Comedy genre models

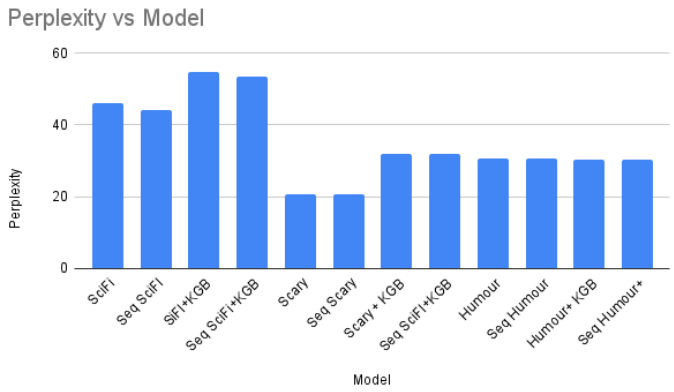


Fig. 9: Perplexity vs Model

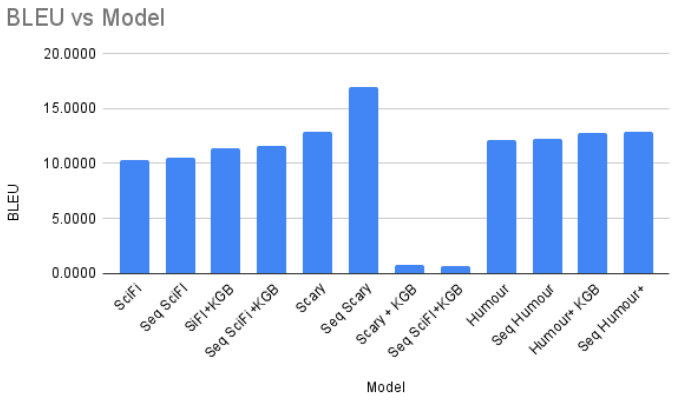


Fig. 10: BLEU vs Model

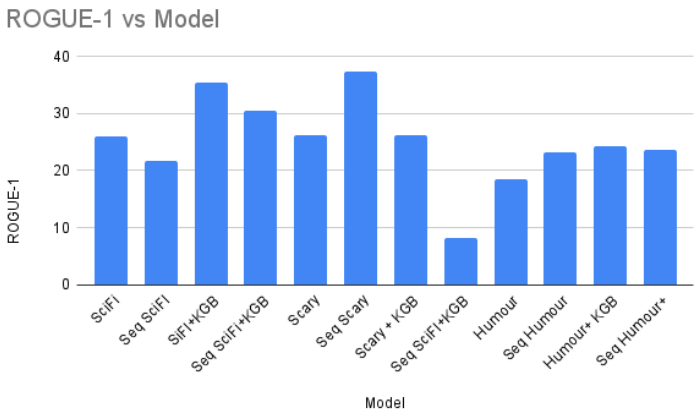


Fig. 11: ROGUE-1 vs Model

Overall, from the above observations from the results there are majorly two takeaways- firstly, the sequential model helps in lowering the perplexity and improving the BLEU and ROUGE scores simultaneously other than this integration of KGB on one side increases the perplexity but also improves the BLEU and ROUGE scores which can be seen in improved quality of stories with better common sense in story lines. The model with sequential generation and KGB integration performs best out of all the models and stands out the most with good stories generated. The results are also confirmed across different datasets to further back the results evaluated. An important observation can be made with reference to the BLEU scores of the Scary + KGB models, as the scores can be seen dropping to extremely low values. This is because of the small size of the Scary stories dataset, and the increase non-contextual word count introduced due to the Knowledgebase. This results in a very small corpus of relevant reference text, hence the low BLEU score.

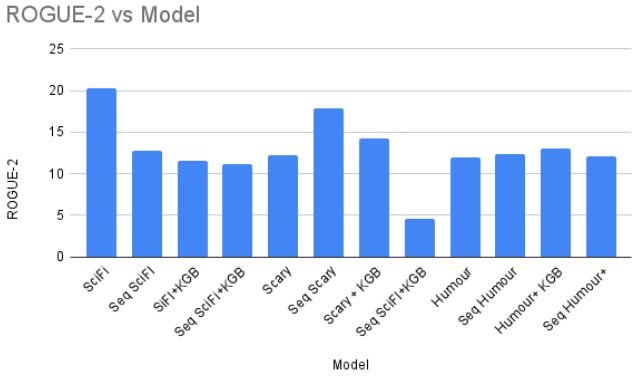


Fig. 12: ROGUE-2 vs Model

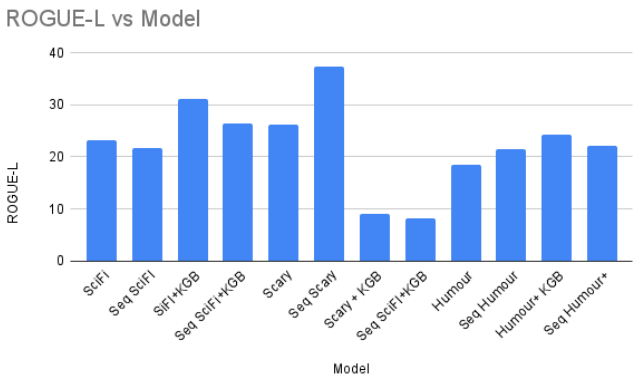


Fig. 13: ROGUE-L vs Model

We identified some key findings of this study. The GPT-2 base model shows a stark inability to maintain context of text over 100 tokens. Since the problem statement also describes that long text is to be generated from a short prompt, problems with information loss have been observed. There is also a marked lack of sentimental and narrative continuation in the stories generated. Experimentation with double fine-tuning and adding commonsense knowledgebase also led to significantly higher perplexity scores, indicating probable reliability issues over the long term. On the positive side, we observe that fine-tuning using a genre specific dataset does result in an effective genre-transformation model, given that the dataset used is large enough. The addition of commonsense knowledgebases results in a much-improved consistency with real-world facts, making the stories much more believable. The sequential implementation delivers a better qualitative result in terms of context management over length of text but cannot be effectively measured through the available metrics.

## IX. Conclusion

The primary goal of this work is to enhance the base GPT-2 model's ability to create stories based on a given prompt. The three main experiments included in the research are double fine-tuning, integration of commonsense knowledgebase, and sequential transformers.

The base GPT2 model works as expected in the task of story generation, with good grammatical precision and accuracy, but lags in terms of artistic expression, creativity, narrative continuity and context, and sentimental continuity. The experiments conducted reveal interesting details about the way GPT-2 responds to modifications such as fine-tuning and knowledge addition.

It is observed that the model's perplexity gets significantly worse upon additional levels of fine-tuning or addition of knowledge, and this can often lead to hallucination and model straying. Another important observation is that adding knowledge base results in a better factual adherence, however creativity is lost to some extent. The performance of Sequential GPT implementations is surprisingly promising despite low scores on the automated metrics. Future research in this domain can be done to develop better metrics and attempt to provide more control over the sentiments in the stories. An important area for future work lies in identifying methods to improve artistic expression in the model as well, which can be done by implementing a conversational interface that allows the user to better control the characters, plot, sentiments, and overall direction of the story. This research enables future researchers to be more informed.

## References

[1] Ammanabrolu, Prithviraj, et al. "Guided neural language generation for automated storytelling." Proceedings of the Second Workshop on Storytelling. 2019.

[2] Chen, Jiaao, Jianshu Chen, and Zhou Yu. "Incorporating structured commonsense knowledge in story completion." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.Jin Chen, Guangyi Xiao, Xu Han, and Hao Chen. Controllable and editable neural story plot generation via control-and-edit transformer. IEEE Access,9:96692–96699, 2021.

[3] Cheong, Yun-Gyung, and R. Michael Young. "Suspenser: A story generation system for suspense." IEEE Transactions on Computational Intelligence and AI in Games 7.1 (2014): 39-52.

[4] Fan, Angela, Mike Lewis, and Yann Dauphin. "Hierarchical neural story generation." arXiv preprint arXiv:1805.04833 (2018).

[5] Ganganwar, Vaishali. "Sentiment analysis of legal emails using Plutchik's Wheel of Emotions in quantified format." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.6 (2021): 4979-4987.

[6] Ganganwar, Vaishali, and R. Rajalakshmi. "Implicit aspect extraction for sentiment analysis: A survey of recent approaches." Procedia Computer Science 165 (2019): 485-491.

[7] Ganganwar, Vaishali, and Ratnavel Rajalakshmi. "Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification." Journal of Information and Telecommunication 8.2 (2024): 167-188.

[8] Ganganwar, Vaishali, and Ratnavel Rajalakshmi. "Enhanced hindi aspect-based sentiment analysis using class balancing approach." International Journal of Information Technology 15.7 (2023): 3527-3532.

[9] Goldfarb-Tarrant, Seraphina, Haining Feng, and Nanyun Peng. "Plan, write, and revise: an interactive system for open-domain story generation." arXiv preprint arXiv:1904.02357 (2019).

[10] Guan, Jian, et al. "A knowledge-enhanced pretraining model for commonsense story generation." Transactions of the Association for Computational Linguistics 8 (2020): 93-108.

[11] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

[12] Mo, Linzhang, et al. "Incorporating sentimental trend into gated mechanism based transformer network for story ending generation." Neurocomputing 453 (2021): 453-464.

[13] Mostafazadeh, Nasrin, et al. "A corpus and cloze evaluation for deeper understanding of commonsense stories." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[14] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[15] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

[16] Rashkin, H., Celikyilmaz, A., Choi, Y., & Gao, J. "PlotMachines: Outline-conditioned generation with dynamic plot state tracking." arXiv preprint arXiv:2004.14967 (2020).

[17] Sap, Maarten, et al. "Recollection versus imagination: Exploring human memory and cognition via neural language models." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.

[18] Saravanan, Shruti, and K. Sudha. "GPT-3 powered system for content generation and transformation." 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT). IEEE, 2022.

[19] Tang, Tianyi, et al. "Mvp: Multi-task supervised pre-training for natural language generation." arXiv preprint arXiv:2206.12131 (2022).

[20] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[21] Wang, Xinpeng, et al. "CHAE: Fine-grained controllable story generation with characters, actions and emotions." arXiv preprint arXiv:2210.05221 (2022).

[22] Wilmot, David, and Frank Keller. "A temporal variational model for story generation." arXiv preprint arXiv:2109.06807 (2021).

[23] Xie, Yuqiang, et al. "Clseg: Contrastive learning of story ending generation." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[24] Yang, Kevin, et al. "Re3: Generating longer stories with recursive reprompting and revision." arXiv preprint arXiv:2210.06774 (2022).

[25] Yao, Lili, et al. "Plan-and-write: Towards better automatic storytelling." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.