Exploring Text and Image Generation from Children's Fairy Tales Using OpenAI Models

Authors: Alex Kramer, Aushee Khamesra, Harishraj Udaya Bhaskar

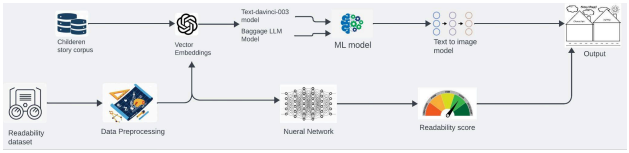
Abstract

This report investigates the intersection of text and image generation through the evaluation of two OpenAI models for the purpose of generating children's fairy tales and corresponding images. The generated text underwent human evaluation and a readability prediction using a neural network model. The study aims to compare the performance of two OpenAI models through human assessment and linguistic analysis.

Introduction

By leveraging two distinct OpenAI models to create narratives and developing corresponding images, we aim to not only showcase the capabilities of these models but also assess their performance through human evaluation and linguistic analysis tools. The following pages dive into the methodology, data, evaluation considerations, models used, results, and discussions of this intersection between evaluating AI through storytelling.

Methodology

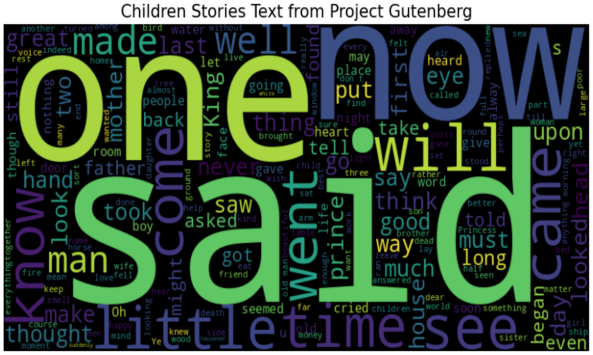


Data

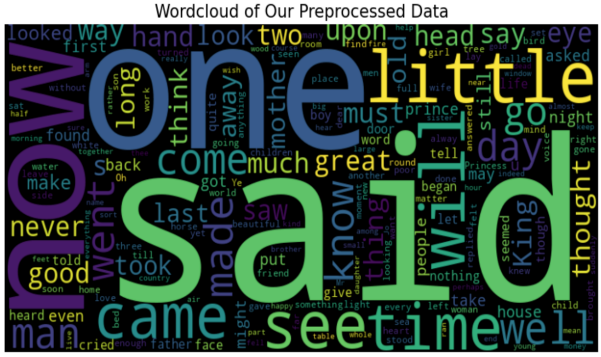
The dataset, comprising 68 children's fairy tales sourced from Kaggle and Project Gutenberg, underwent rigorous cleaning. Two OpenAI models, text-davinci-003 and babbage, were selected for text generation based on their natural language processing proficiency. Preprocessing involved story segmentation into 'chunks,' randomly selecting 1,500 chunks for diversity. Image generation employed the AutoPipelineForText2Image from Diffusers, pretrained on stable-diffusion-xl-base-1.0. LangChain facilitated creative story generation, summarization with Transformers provided concise prompts, and the text2image pipeline translated prompts into visually appealing images.

Exploration Analysis

Children Stories Text from Project Gutenberg



Wordcloud of Our Preprocessed Data



In our exploratory analysis, we conducted a comparison of word clouds derived from both the original and preprocessed datasets to gain insights into the impact of data preprocessing. Surprisingly, the word clouds exhibited remarkable similarity, suggesting that our preprocessing steps effectively retained the fundamental linguistic characteristics present in the original data.

Models Working Mechanisms

Text & Image Generation Models

Utilizing LangChain, OpenAI's text-davinci-003 and babbage models generated imaginative children's stories based on user input and document queries. The GPT-3 models exhibited a nuanced understanding of language, ensuring coherent and contextually relevant narratives. The AutoPipelineForText2Image, pretrained on stable-diffusion-xl-base-1.0, facilitated the generation of visually compelling images from the text generated by OpenAI models.

OpenAI models, text-davinci-003 and babbage, demonstrated proficiency through creative text generation. Trained on diverse internet text, they

employed advanced decoding algorithms, often based on beam search, to generate contextually relevant text sequences. Handling unknown words involved leveraging contextual information and linguistic structures, highlighting the robustness of OpenAI language models.

Readability Model

The readability score model underwent multi-step preprocessing, including tokenization, stopword removal, and lemmatization. Features such as sentence length, word count, and lemma length, along with counts of punctuation and quotes, were extracted. TF-IDF vectorization and Count Vectorization transformed text into numerical representations. The neural network, implemented using TensorFlow and Keras, featured multiple dense layers. Trained on mean squared error, the model predicts readability scores, offering a comprehensive assessment of text difficulty.

Evaluation Criteria

Human Centric Evaluation

To assess the quality of generated stories, a comprehensive human survey was developed. Participants were presented with three sets of stories for ranking: one generated by the babbage model, one by the text-davinci-003 model, and one crafted by a human. In total, survey takers evaluated and ranked nine stories, ordering them from best to worst. The evaluation criteria encompassed grammaticality, kid-friendliness, and narrative flow. This human-centric approach provided valuable insights into the perceived quality of stories, considering aspects crucial for children's engagement.

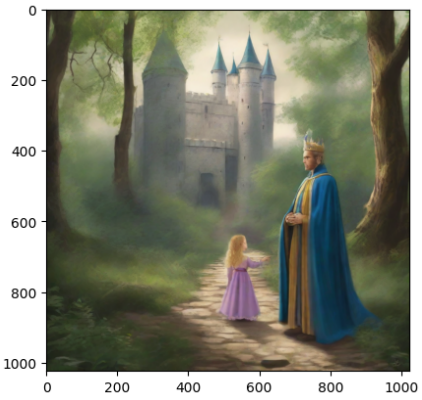
Readability Score Evaluation

A neural network model designed to predict the difficulty of a text through the analysis of words, sentence length, and semantics was employed for the readability score evaluation. Adapted from a Kaggle GitHub repository, the readability score ranged from -1 (most complex/unreadable) to 1 (most readable). However, before inputting the generated text into the readability model, a series of preprocessing steps were undertaken. The custom preprocessing functions collectively transformed raw text data into a machine-learning-friendly format.

Results

Example

The inclusion of a specific text and image generation example serves as a tangible illustration of the capabilities of the OpenAI models in crafting imaginative narratives and translating them into visuals. The example below provides readers with an understanding into what the model is generating. This example was generated by text-davinci-003.



“Once upon a time there lived a beautiful Princess who was beloved by everyone in her kingdom. She had long, golden hair and eyes like sapphires. Everyone adored her, from the youngest of children to the oldest of elders.One day, the Princess heard of a magical kingdom far away, in a distant land. She decided to set off on a journey to this kingdom, to explore its wonders and meet its people.The Princess packed her things and set off on her journey…”

Human Survey Results

The human-centric evaluation of the generated children's fairy tales engaged a cohort of 10 participants, assigned with the task of ranking stories based on grammaticality, kid-friendliness, and narrative flow. Each participant meticulously evaluated three sets of stories: those generated by the babbage model, the text-davinci-003 model, and narratives crafted by a human author. The aggregated results, graphically depicted in Figure 1.1, illuminate the perceived quality of the generated stories. Notably, the text-davinci-003 model excelled in grammar and flow, surpassing both the human-written stories and the babbage model.

However, in terms of kid-friendliness, the human-authored story emerged as the highest scorer.

Figure 1.2 provides a comprehensive overview, illustrating that the text-davinci-003 model achieved the highest scores across all evaluation categories, with human-written stories closely trailing. In a noteworthy observation, Figure 1.3 reveals that the human-authored stories outperformed both OpenAI models across all categories. A distinctive finding is the babbage model's superior performance in grammar compared to the text-davinci-003 model, adding an intriguing layer to the evaluation outcomes. These graphical representations offer detailed insights into the strengths and weaknesses of each model, contributing to a nuanced understanding of their performance in crafting children's narratives.

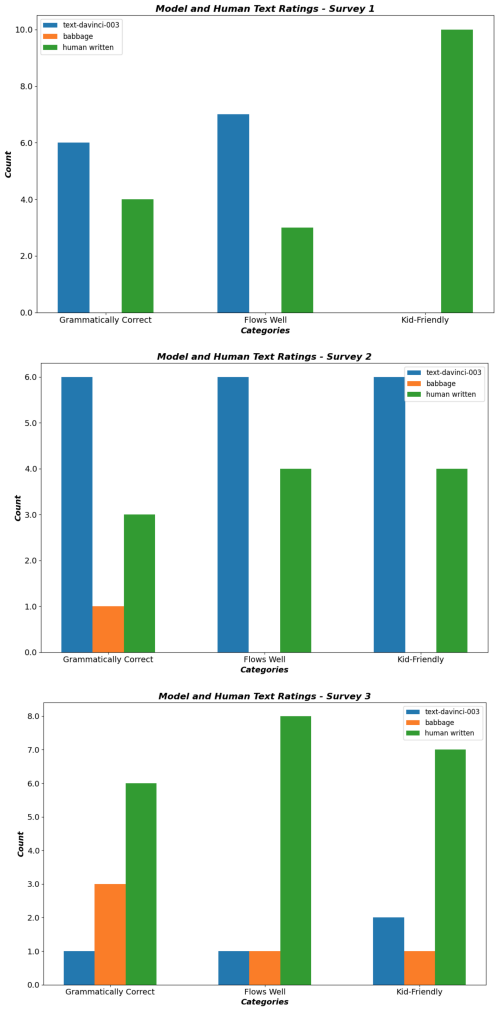
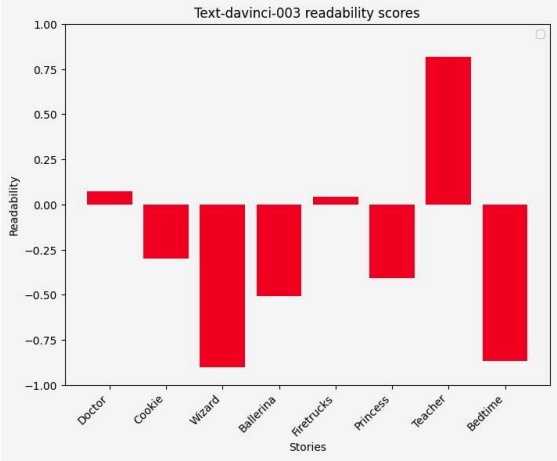


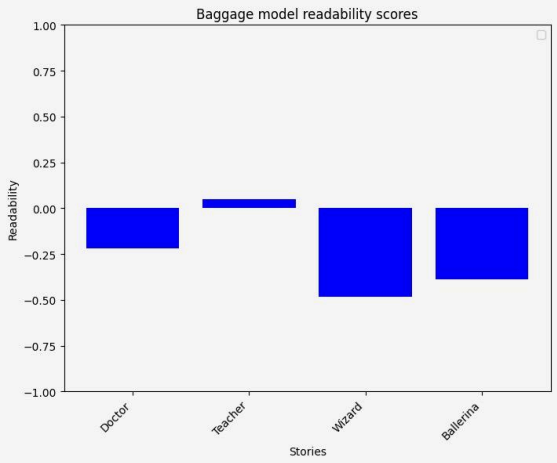
Figure 1: a) Story 1 Results b) Story 2 Results c) Story 3 Results

Readability Score

The evaluation of readability scores added an insightful dimension to our study, offering an objective assessment of the generated stories. Notably, the text-davinci-003 model exhibited commendable performance, with three of its generated stories achieving positive readability scores. In contrast, the babbage model produced one positively scored story, highlighting a nuanced difference in the models' abilities to create text with varying degrees of complexity and readability. Human authors produced one story that received a positive readability score, showcasing the model's competition with human-generated content in terms of readability.

When generating stories using the babbage model, certain themes posed challenges, attributed to a lack of data on those specific topics. For instance, themes such as fire trucks were absent in the training data, causing the babbage model to struggle in generating corresponding stories. This limitation underscores the significance of diverse and comprehensive training data for ensuring the model's proficiency across a wide array of themes and narrative elements. In contrast, the text-davinci-003 model showcased greater adaptability, successfully generating stories even in the absence of specific theme-related data.





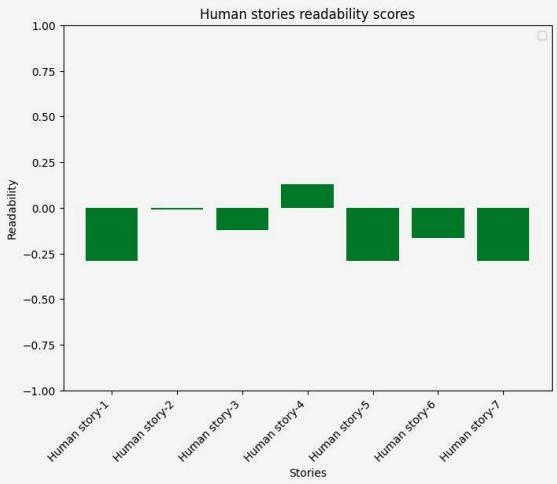


Figure 2: a) Text-davinci-003 readability scores
b) Baggage model readability score*
c) Human authored readability scores

Conclusions

The human-centric evaluation, involving a cohort of 10 participants, revealed nuanced aspects of model performance. While the text-davinci-003 model exhibited superiority in grammar and flow, the babbage model showcased unexpected strength in grammar, adding complexity to our understanding of their respective strengths and weaknesses. Unsurprisingly the human-authored stories consistently outperformed both OpenAI models across all categories, emphasizing the importance of human creativity and intuition in crafting engaging children's narratives.

In the realm of readability scores, the text-davinci-003 model proved adept, achieving positive scores for three generated stories. The babbage model, while generating positively scored stories, faced challenges with certain themes due to limited training data. The comparison with human-authored stories in readability scores indicated a competitive landscape, highlighting the ongoing need for refining AI models to match or surpass human-authored content.

Noteworthy conclusions include the importance of a diverse and comprehensive dataset for model proficiency, as evidenced by the babbage model's struggles with theme-specific narratives. The study also suggests that the best-performing AI model can generate stories on par with human-authored ones when given appropriate training data. However, the presence of a "bad" model, indicated by the babbage model's outlier in grammatical scores, emphasizes the need for continuous adjustments and refinements, particularly in the context of the readability model.

Future Improvements

Looking forward, our team would like to further develop several areas of this project in order to create a more meaningful and holistic analysis of these different types of models. Expanding the training dataset with a more diverse collection of children's stories, and with increased computing power, promises more sophisticated model architectures and efficient handling of larger datasets. The consideration of training models from scratch, particularly emphasizing on children's literature, offers a specialized learning approach. Generating and evaluating a broader set of stories enhances the project's insights, while expanding its

application to complex texts and chapter-wise image generation enriches the narrative experience. We would like to perform an exploration into the factors that impact the Readability model to uncover potential biases, in order to foster a more equitable assessment of stories.

Work Cited

- https://www.linguisticanalysistools.org/arte.html?ref=commonlit.org
- https://www.kaggle.com/competitions/commonlitreadabilityprize/data
- https://platform.openai.com/docs/models/overview
- https://www.linguisticanalysistools.org/arte.html?ref=commonlit.org
- https://www.youtube.com/watch?si=RTMQbxS2wcixLR21&v=lG7Uxts9SXs&feature=youtu.be
- https://platform.openai.com/docs/models/gpt-3-5
- https://blamouche.medium.com/a-quick-openais-language-models-comparison-9987ddb2a723
- https://github.com/suhasmaddali/Predicting-Readability-of-Texts-Using-Machine-Learning/blob/main/Predicting%20Readability%20of%20Texts%20Using%20Machine%20Learning.ipynb