UNIVERSITY OF CALIFORNIA
SANTA CRUZ

STABILITY.GPT: COMBINING GPT-4 AND STABLE DIFFUSION TO GENERATE STORYBOOKS THAT ARE TEXTUALLY AND VISUALLY COHESIVE

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTATIONAL MEDIA

by

Anirudh Venkataramanan

December 2023

The Thesis of Anirudh Venkataramanan
is approved:

Professor Jim Whitehead, Chair

Professor David Lee

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

stability.gpt: Combining GPT-4 and Stable Diffusion to Generate Storybooks that are Textually and Visually Cohesive

by

Anirudh Venkataramanan

Generative artificial intelligence is strong at creative tasks. Image generation models like Stable Diffusion, Dall-E, and Midjourney generate images given a prompt. Language learning models like ChatGPT can generate text. Language learning models' strengths are generating creative content like stories and poems. While current systems that combine the power of language learning models and image generation models to create storybooks exist, they lack image consistency - both in terms of style and visually depicting the text on the page.

This paper introduces stability.gpt - A new modern system using the latest version of GPT (GPT-4) and Stable Diffusion to generate storybooks that have consistency in image style. Upon having some users create stories with stability.gpt and analyzing ten stories from a diverse set of genres, it becomes clear that while character consistency has improved and that the images support the story text, there is still room for improvement in generating consistent animals, objects, and styles.

Acknowledgments

I want to thank my faculty advisor Dr. Jim Whitehead for his support and encouragement as I worked on this paper. I also want to acknowledge all the students in the Augmented Design Lab at the University of California, Santa Cruz for their companionship and advice.

Chapter 1

Introduction

Generative artificial intelligence (AI) is a lucrative technology whose market is expected to expand in the coming years, currently being predicted to amass $63 billion US dollars by 2028 [20]. Language learning AI models (LLMs), such as OpenAI's GPT-4, Google's Bard, and Meta's Llama 2, are currently used to generate textual content whereas image generation AI models like Dall-E, Stable Diffusion, and Midjourney are used to generate images. Each generative AI model is trained on and learns from large datasets [39]. Bard is a LLM that is connected to the internet and constantly learning, thus making it up-to-date on the latest information. [15]. On the other hand, GPT-4 and Llama cannot access datasets and have a knowledge cutoff of Septembebr 2021 and September 2022, respectively. Both models are also incapable of learning from their experience. [22][34].

LLMs are nondeterministic by default. This means they can give different answers to the same questions at different points in time, with some answers being incorrect or inaccurate [43]. Therefore, it is crucial to take responses from LLMs with a grain of salt, especially when asking for factual information or solutions to technical problems. In spite of their nondeterminism, LLMs have proven to be powerful tools with vast capabilities, and ChatGPT - an online chatbot that uses GPT-3.5 and GPT-4 when users upgrade to ChatGPT Plus - has over 100 million users as of July 13, 2023 [18]. The rest of this paper will focus on ChatGPT and the GPT LLMs specifically given its large user base - students and working professionals alike - all around the world as well as its relevance to my research study.

GPT LLMs have a temperature parameter ranging between 0 and 1 that affects how focused and deterministic outputs are. Lower temperatures generally result in more deterministic responses from the LLM that are in line with what the prompt is requesting. On the other hand, larger temperatures result in more creative responses [9][31][42][43]. While determinism does mean that the same output is likely to be provided when given a particular input multiple times, it does not guarantee that the output will be reliable.

LLMs are more strong in creative tasks where there isn't one (or a few) objectively right answers. LLMs are capable of generating all sorts of creative content, such as raps, poems, jokes, prompts for image generation, haikus, and more. However, ChatGPT and most LLMs cannot generate illegal or violent content as per their content policy [59]. Some creative content can also be multimodal, combining text and images. This includes storybooks, comic books, slideshows, and magazine articles [33]. LLMs are not capable of generating visual media on their own. Adobe has announced that their image generation model, Adobe Firefly, will be integrated into Google Bard in the future, but that is not available yet [23]. ChatGPT has a plugin from Argil AI that lets users generate images, but it is only available for ChatGPT Plus users who pay $20 a month, making ChatGPT an expensive tool for generating images [13]. Furthermore, the plugin uses Dall-E and not a fine-tuned stable diffusion image generation model. This negatively impacts the generated image quality [8]. Argil AI has their own studio that lets users create their own AI workflows. It is possible to create a workflow that generates a story via ChatGPT and then generates images for that story with stable diffusion. However, these workflows do not take any user feedback after each step and make any changes. They run once on their own. Users would have to make stories on their own or run the entire workflow again, doubling the costs to create a story [2]. Even outside of a cost perspective, it should be noted that the characters and styles across multiple images can be different, as shown in Figure 1.1.

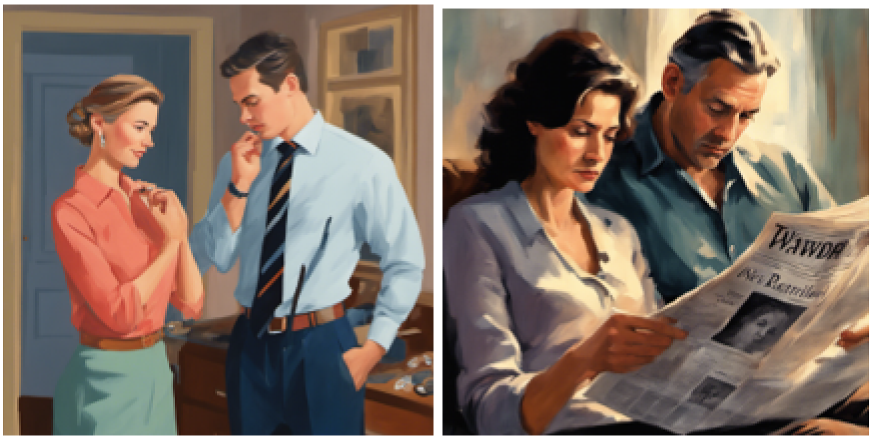
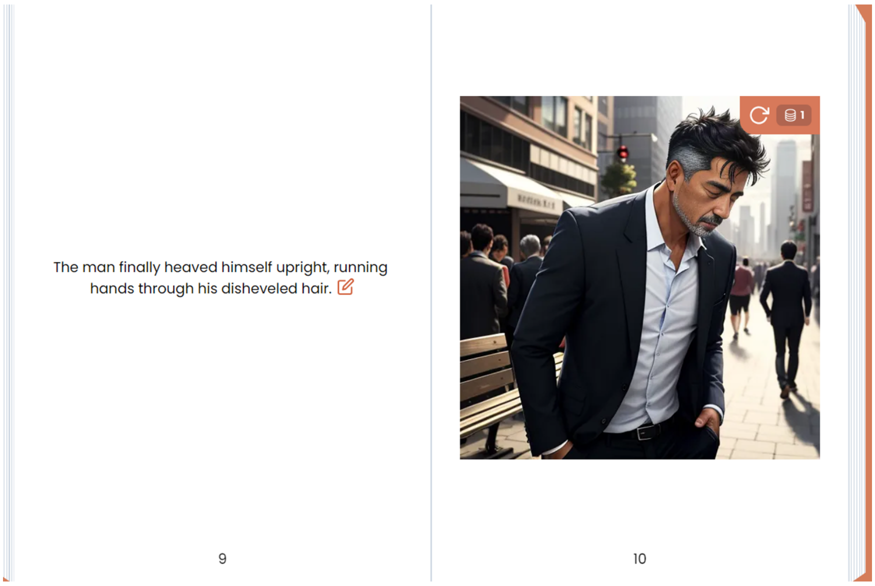


Figure 1.1: Two images generated from Argil AI for a story of a husband and wife getting ready for work in the morning. The color and style of their hair differs between both images as does the clothes they are wearing.

StoryBird is another cheaper alternative to creating multimodal content, particularly stories, with AI [27]. StoryBird is easier to use as all the user needs to do is input a story synopsis. The story will then be generated with GPT 3.5 and Midjourney [58]. The issue that yet again arises is the lack of consistent characters in all generated images. Users can re-generate images and edit parts of the story, but the occasional disconnection between the story and corresponding images as well as the consistency of the images is apparent as shown in Figures 1.2 and 1.3.



The man finally heaved himself upright, running hands through his disheveled hair.

Figure 1.2: Disconnect between the story text and the corresponding image in StoryBird. In the story, the man has just woken up and is still in bed. However, in the image, he is all dressed up and outside.

The lack of consistent characters and disconnect between text and images are apparent in existing AI multimodal content generators. In order to make artificial intelligence a more lucrative tool for visual storytelling, research questions that will be explored in depth in this paper are as follows -

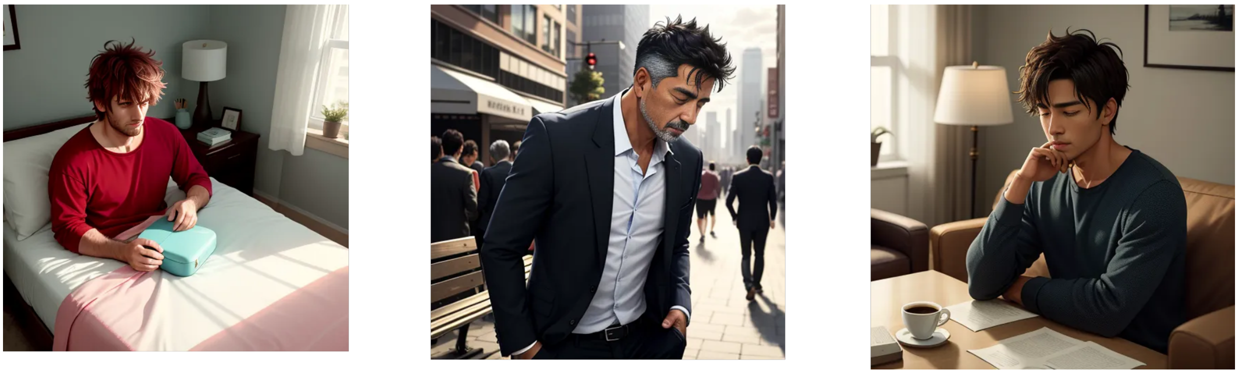


Figure 1.3: The same character in a StoryBird story looking different across multiple images.

1. How can prompts for image generation AI models - specifically stable diffusion - be engineered to maintain consistent styles, locations, characters, and objects across multiple images?
2. How can GPT-4 and Stable Diffusion work together to create picture storybooks that have consistency across all images?
3. For which genres can GPT-4 and Stable Diffusion create storybooks that are the most accessible and coherent both literally and visually?

To answer these questions, I created an AI storybook generator called stability.gpt that creates storybooks that have images with consistency and a variety of frames/perspectives based on the context of the story. GPT-4 and Stable Diffusion both have publicly available official APIs, which is why I chose to use them to create this system [40][56]. Chapter 3 further discusses my choice of LLM and image generation model in developing this system. By creating and evaluating stability.gpt, I learned about how to construct stable diffusion prompts to optimize image consistency and the best ways to use GPT-4 to automatically construct the right prompts for stories.

Chapter 2 takes a look at how GPT models are used in storytelling, existing methods of generating images with a consistent style for visual narratives, the best image generation model to use to generate faces, and how GPT-4's prompts can be engineered to follow specific instructions. The development and evaluation of stability.gpt are discussed in the subsequent chapters.

Chapter 2

Literature Review

In this section, the use of LLMs for storytelling, prompt engineering with LLMs to receive a specific output, a comparison of image generation models, and strategies to maintain consistency across multiple AI generated images will be discussed.

2.1 Collaborative Storytelling with LLMs

Collaborative storytelling is a common use of LLMs. This is when humans and AI take turns adding sentences to a story [38][52]. SAGA is a web application that lets multiple people and GPT-3 collectively add to a story one at a time. Since neither the human writers nor the AI have planned out the plot and structure of the story ahead of time and are building off of each other, plot twists added by one author will be a surprise for the others, making collaborative storytelling exciting and fun. Stories in SAGA are more immersive when the main characters are the human authors themselves [52]. A collaborative storytelling study from Eric Nichols et. al. shows that LLMs ranked and tuned on storytelling data provide more interesting, engaging, and human-like story contributions [38].

Some limitations of collaborative storytelling are that the story may lack cohesion depending on what the humans and AI add to the story given the previous contributions. The human-preferred tone and mood of the story may also not be met [38]. From this, it is clear that it is best when either AI or humans are the sole author of a story and there is no cross-collaboration. The non-author(s) can provide the premise of the story and give feedback as it is being written but having one voice behind the story will ultimately lead to a story that flows better.

2.2 Prompt Engineering for LLMs

It is evident that LLMs are capable of generating creative text content. However, sometimes users may want the generated output to be structured a certain way or to contain some crucial information. This can be accomplished by prompt engineering - tailoring prompts to get the LLM to give the consumer the output they want [47][53].

In order to learn about the role of AI in entrepreneurship, Cole E. Short et. al. asked ChatGPT to mimic different celebrities' writing styles and fine-tuned their inputs to get their desired output. For four different CEOs - Elon Musk, Indra Nooyi, Tony Hsieh, and Lisa Su - the researchers first asked ChatGPT the following prompt - *Write an elevator pitch in the style of [CEO Name].* For Musk, ChatGPT's response showed the sensibilities of a creator. They then fine-tuned their prompt to ask ChatGPT to rewrite its elevator pitches with store storytelling. The resulting pitches were longer and had three new components found in stories - a subject that is looking for something, the reason they are looking for that thing, and forces that negatively impact the subject. Subsequently, ChatGPT was asked to write crowdfunding pitches and Twitter pitches, and then to rewrite those pitches with storytelling for the same four celebrities. By specifying the type of pitch they wanted, Cole et. al. were able to get pitches addressed to different audiences and in a style suitable to each audience. For example, tweets have a limited character count so naturally the Twitter pitches were shorter than the others. The Twitter pitches even contained hashtags, a tool used to group similar tweets together [53].

Ultimately, AI and LLMs do not know everything. Their knowledge is limited to the datasets they have been trained on and the resources they can access. For example, ChatGPT is not designed for screenwriting. The scripts it produces may not follow the correct format or deliver a story that is coherent [44]. One way to maximize the chances of receiving an ideal output is to provide examples to the LLM. This is especially true when asking the LLM to generate something that follows a specific format, such as prompts for image generation models like Stable Diffusion, Dall-E, and Midjourney [4].

Sometimes, users may want an LLM to accomplish something big. It is advised to split up a big task into smaller subtasks. This will result in a lower error rate [41]. For instance, users could ask ChatGPT to generate a story in one prompt and to generate prompts for an image generation model in another prompt rather than instructing it to do both in one prompt.

2.3 Dall-E vs. Midjourney vs. Stable Diffusion

A study conducted by Ali Borg from Quintic AI that quantitatively compared Dall-E, Midjourney, and Stable Diffusion revealed that Stable Diffusion generates the most realistic faces. This conclusion was reached by comparing the AI-generated faces to real faces using the Fréchet Inception Distance (FID) scale. Lower FID scores correspond to better generated faces and Stable Diffusion images had the lowest FID scores [10].

While there is evidence that Stable Diffusion generates the best faces, there is no concrete evidence on which image generation model generates the best images overall. What model is best is subjective and based on each individual's needs and preferences. Unlike Midjourney, Stable Diffusion and Dall-E both have official public APIs that can be used to build software applications [30][49]. Between Stable Diffusion and Dall-E, Stable Diffusion is the better option for developers as it is cheaper to use and gives users much more options to customize their output [24].

2.4 Consistent Styles Across Multiple Images

Prompt engineering is crucial with image generation models to get the desired visual output. Adjectives have a smaller impact on the generated image as they are describing the cosmetics of nouns. Nouns, on the other hand, are introducing new places, people, or objects and influence the generated image a lot more. Mentioning the name of an artist whose style the image should mimic can completely transform how an image is structured [64]. Bad image prompts are usually very short, not in English, and/or contain emoticons. Different models follow different prompt syntax. Stable diffusion uses commas to separate keywords unlike its contemporaries in Midjourney and Dall-E [62].

The keywords man and woman in a prompt are ambiguous and can be interpreted in many different ways by an image generation model every single time. On the other hand, including the name of a celebrity is much more likely to result in a character being generated consistently across all images as long as the physical descriptors - such as hair color, outfits, and overall style - also remain the same. Prompts can also include the names of multiple celebrities for one character. The prompt *[Brittany Spears — Vanessa Hudgens]* will result in a character that looks like a mashup of both Brittany Spears and Vanessa Hudgens [16].

The classifier-free guidance (CFG) scale in Stable Diffusion affects how much a generated image adheres to the prompt. Negative CFG values result in images completely different from the prompt [62]. In general, larger CFG values allow stable diffusion to follow the prompt more closely at the cost of quality whereas lower CFG values gives stable diffusion far more creative control and results in better image quality. Experiments with CFG values have shown that a value around 7 is a good balance between creative freedom and prompt cohesion [28][48].

Many image generation models including Stable Diffusion also have a seed. Seed determines the random noise that Stable Diffusion will then denoise to produce the final image output [35]. This is why the exact same prompt and Stable Diffusion settings with the same seed will always produce the same output whereas a different seed changes the image that is generated. Having a fixed seed will ensure that all generated images have a similar composition [17].

Denoising in Stable Diffusion is known as sampling or scheduling. The number of inference steps impacts how much noise is removed during each step of the denoising process. There are various samplers/schedulers which control the noise levels during the image generation process in their own way and impact the image quality. Some are slower than others but also prioritize higher image quality [6][36].

Chapter 3

stability.gpt

Stability.gpt is an online system for generating storybooks. Users first enter a story synopsis (and optionally, a story title and the number of pages the book should have).. This information is fed into GPT-4 to generate text for the story. GPT-4 then automatically generates Stable Diffusion image prompts for each page of the story. The prompts are generated with image style consistency as the goal. These image prompts are then passed into Stable Diffusion and the images are generated.

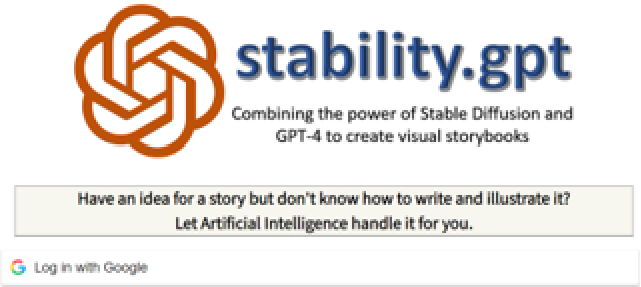
This section will provide an overview and then a deeper look at using stability.gpt by explaining in detail the various web pages the user will encounter while using the system and also what will be happening behind the scenes in the backend to generate the content that the user sees.

3.1 Overview of stability.gpt

Figure 3.1 shows the homepage of stability.gpt when the user is not logged in and Figure 3.2 shows the homepage when the user is logged in. Users can login with their Google account. All students at my university - the University of California, Santa Cruz - are given Google accounts and Gmail is the most popular email provider which is why I chose this method of authentication [21][29].

As the headline of the stability.gpt logo in Figures 3.1 and 3.2 state, stability.gpt uses GPT-4 - the latest version of GPT from OpenAI - and Stable Diffusion to generate visual storybooks. This is accomplished by using the APIs for both AI models. Both AI models work with each other to create storybooks with consistent and appropriate images.

Upon clicking the *Go to Dashboard* button in Figure 3.2, users will be brought to the page shown in Figure 3.3. Here, users can choose to either read books they have already generated or create a new storybook.

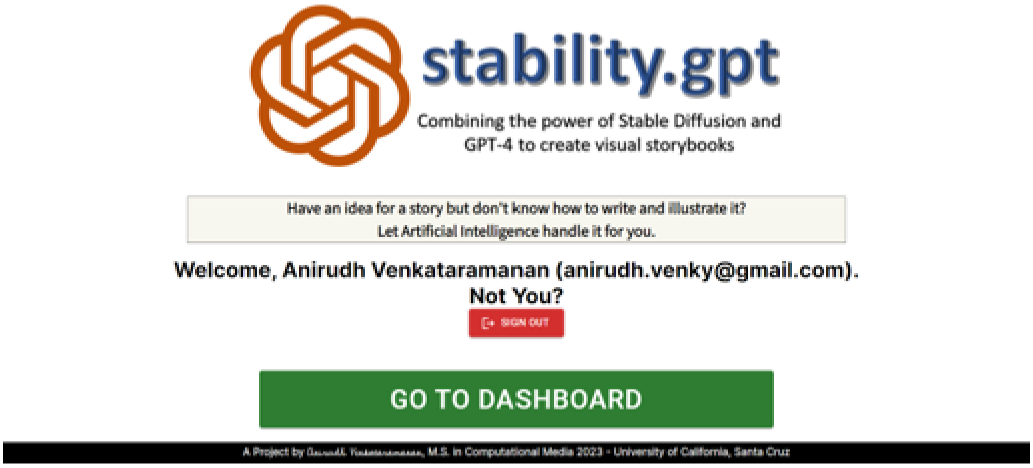


stability.gpt

Combining the power of Stable Diffusion and GPT-4 to create visual storybooks

Have an idea for a story but don't know how to write and illustrate it? Let Artificial Intelligence handle it for you.

Figure 3.1: Homepage of stability.gpt when the user is not logged in.



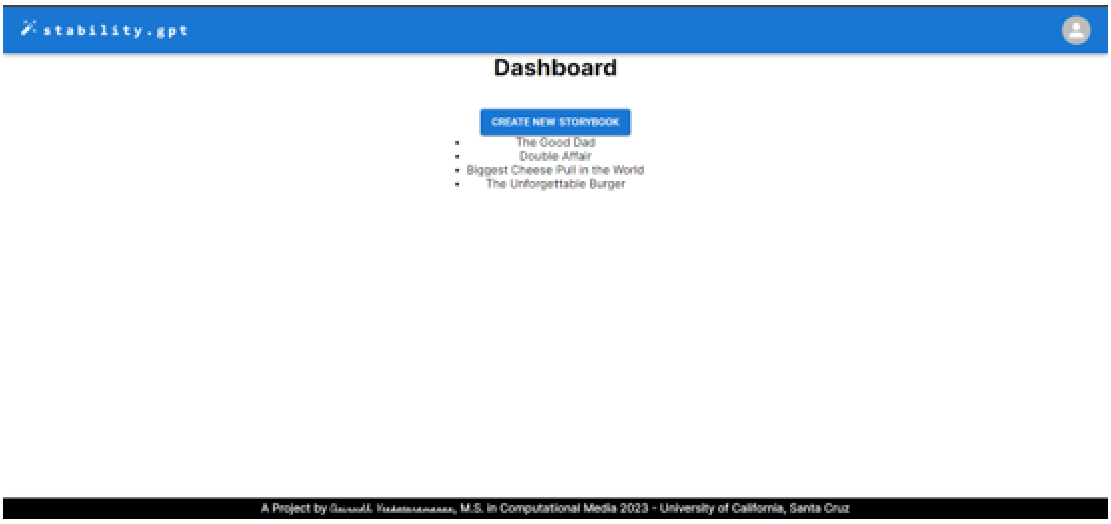
stability.gpt
Combining the power of Stable Diffusion and GPT-4 to create visual storybooks
SIGN OUT
GO TO DASHBOARD

Have an idea for a story but don't know how to write and illustrate it?
Let Artificial Intelligence handle it for you.

Welcome, Anirudh Venkataramanan (anirudh.venky@gmail.com). Not You?

A Project by Anirudh Venkataramanan, M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.2: Homepage of stability.gpt when the user is logged in.

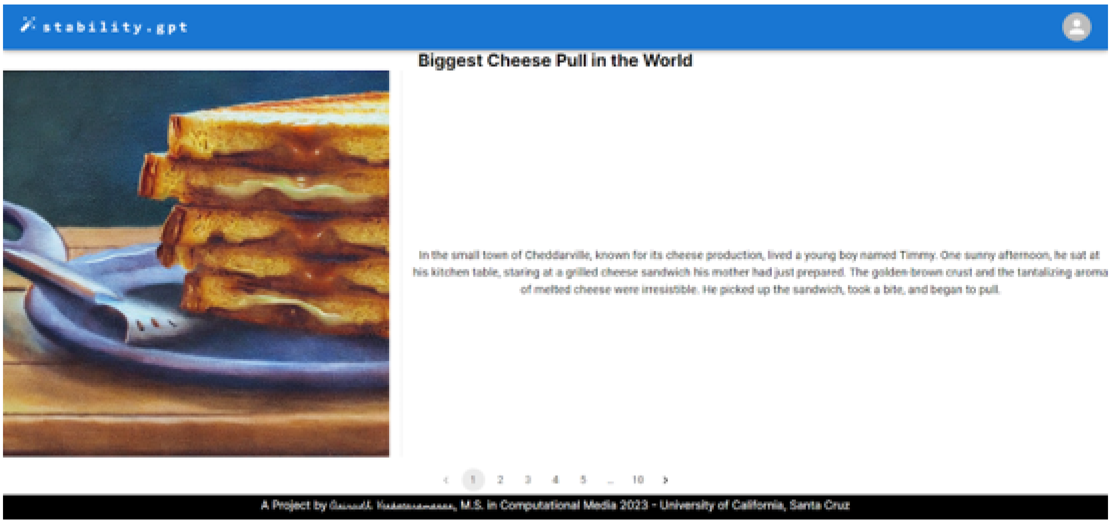


- The Good Dad
- Double Affair
- Biggest Cheese Pull in the World
- The Unforgettable Burger

A Project by Anirudh Venkataramanan, M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.3: stability.gpt dashboard where users can either select a book they have already generated to read or choose to create a new storybook.

Figure 3.4 shows the user interface of a book generated from stability.gpt. The title of the book is in bold text at the top of the screen. The book is divided into pages that can be accessed from the pagination bar at the bottom of the screen. The image for each page is to the left of the screen and the text is to the right. Both the text and images are divided by a light vertical line.



Biggest Cheese Pull in the World

In the small town of Cheddarville, known for its cheese production, lived a young boy named Timmy. One sunny afternoon, he sat at his kitchen table, staring at a grilled cheese sandwich his mother had just prepared. The golden-brown crust and the tantalizing aroma of melted cheese were irresistible. He picked up the sandwich, took a bite, and began to pull.

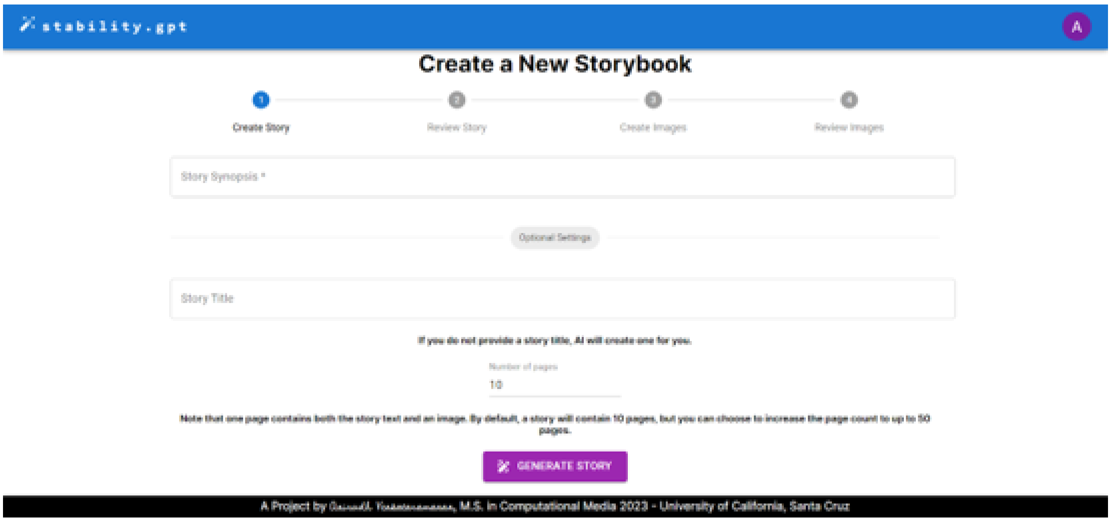
A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.4: User interface of a book generated from stability.gpt. The title is in bold text at the top of the screen, the image and the text take up the most space, and a pagination bar is shown at the bottom to allow users to easily access different pages of the book.

The rest of this section will describe how storybooks are generated with stability.gpt - what information the user provides and what GPT-4 and stable diffusion does with that information to create the story that meets the consumer's needs.

3.2 Generating story text with GPT-4

To create a storybook, users must click the *Create new Storybook* button from the stability.gpt dashboard shown in Figure 3.3. Upon doing so, they will be brought to the *Create a New Storybook* page shown in Figure 3.5.



Create a New Storybook

Story Synopsis *

Story Title

If you do not provide a story title, AI will create one for you.

Number of pages: 10

Note that one page contains both the story text and an image. By default, a story will contain 10 pages, but you can choose to increase the page count to up to 50 pages.

A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.5: First step of the *Create a New Storybook* page in stability.gpt where users provide details on the story plot.

Creating a storybook with stability.gpt is split into four tasks as recommended by the OpenAI documentation [41]. The first step in the *Create a New Storybook page* is describing what the story is about. Like most AI models, GPT-4 and Stable Diffusion are not capable of generating inappropriate not-safe-for-work (NSFW) content [12]. Therefore, users are currently only allowed to generate stories without violent or explicit content.

Aside from the aforementioned two restrictions, users can create stories about practically anything. Story titles and page numbers can be optionally provided by the user. A title will be generated by GPT-4 if it is not provided. By default and at minimum, each storybook will contain 10 pages. A maximum of 50 pages is set to keep storybooks from being too long. Most picture books contain 32 pages [25]. As each page in my storybook is divided into two parts and contains both text and an image, 10 pages is equivalent to 20 pages in a printed book. I did not make the minimum page count any lower as I was concerned that stories would not have enough content with very few pages.

Upon clicking the *Generate Story* button, users will see the loading page in Figure 3.6 as the story text gets generated from the GPT-4 API.

Generating...

Figure 3.6: Loading page that users are shown as the storybook is being generated.

GPT-4 works behind the scenes while the user sees the loading screen. The following prompt is sent to GPT-4, with content in parentheses being things that are passed in from the form in Figure 3.5 -

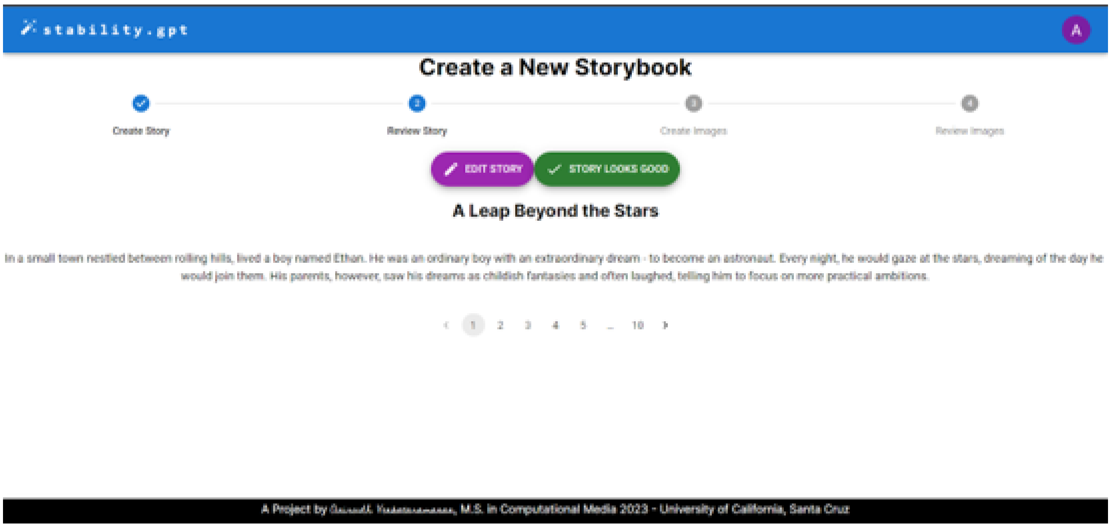
Generate a story with the prompt mentioned below. Separate the story into (number of pages specified) different pages and label the start of each page with "PAGE #" where "#" is replaced by the page number. Do not include any bonus pages. If you think the page count is too high, add fewer sentences per page. Make sure you extend the story to fit all pages. Each page should legitimately push the story forward. Write *the story from a third person point of view. At the end of the story, write the text "THE END". (If the user has not specified a title, the following line will be appended to the prompt - At the start of the story, before the first page, enter a story title preceded by the text "TITLE:".)*

The story prompt is as follows - (story synopsis from the user)

Many of the elements of the prompt are present due to GPT-4 initially generating things I did not want such as first-person narratives and bonus pages. Specifying that these things are prohibited prevents GPT-4 from generating them. I also stated that each page should be preceded by the text *PAGE #*. That way, when the response from GPT-4 is generated, it will be easy to parse out the text for each page and display it to the user. From my testing, I also realized that a lower temperature increased the likelihood of GPT-4 following my instructions and not hallucinating. With the lowest temperature of 0, I have never had an invalid response returned from GPT-4.

Once the story is generated, users will be brought to the second step of the *Create a New Storybook* page where they can read the story that has been generated with GPT-4. There is a pagination bar towards the bottom of the screen for them to flip pages. This is shown in Figure 3.7.

If the user likes the story they read, they can click the *Story Looks Good* button. Otherwise, they can click the *Edit Story* button. Upon doing so, a window will slide in from the right of the screen taking up half of the screen (or the entire screen if stability.gpt is being accessed from a smartphone device) as shown in Figure 3.8. This new window contains a textbox where users can describe what changes they want made

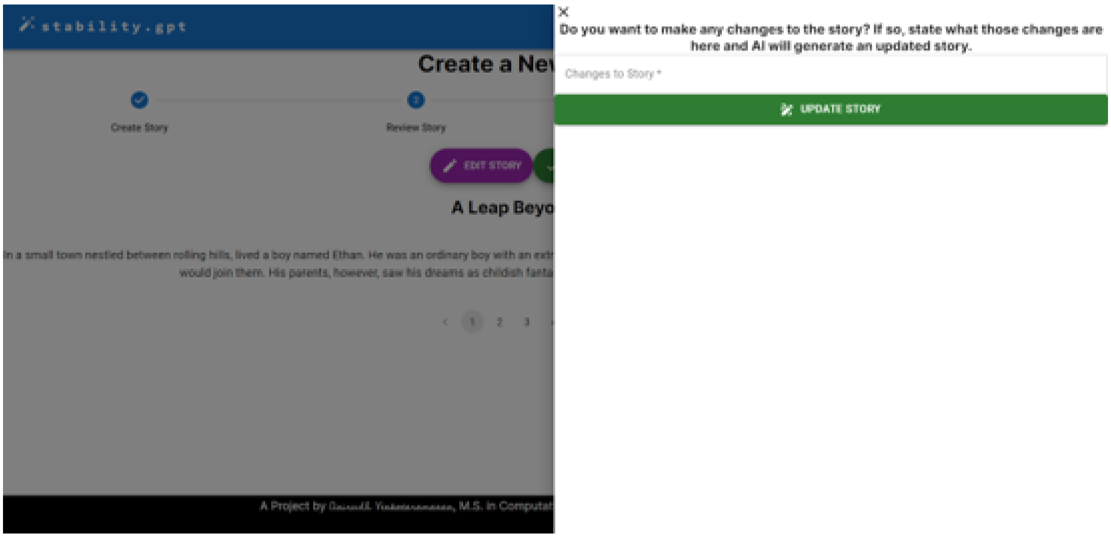


In a small town nestled between rolling hills, lived a boy named Ethan. He was an ordinary boy with an extraordinary dream - to become an astronaut. Every night, he would gaze at the stars, dreaming of the day he would join them. His parents, however, saw his dreams as childish fantasies and often laughed, telling him to focus on more practical ambitions.

A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.7: The second step of creating a storybook in stability.gpt is *Review Story*. Users can read the story that GPT-4 generated and affirm that it is good or suggest changes.

to the story. They can even choose to change the story title or the number of pages the story has.



In a small town nestled between rolling hills, lived a boy named Ethan. He was an ordinary boy with an extr would join them. His parents, however, saw his dreams as childish fanta

Do you want to make any changes to the story? If so, state what those changes are here and AI will generate an updated story.

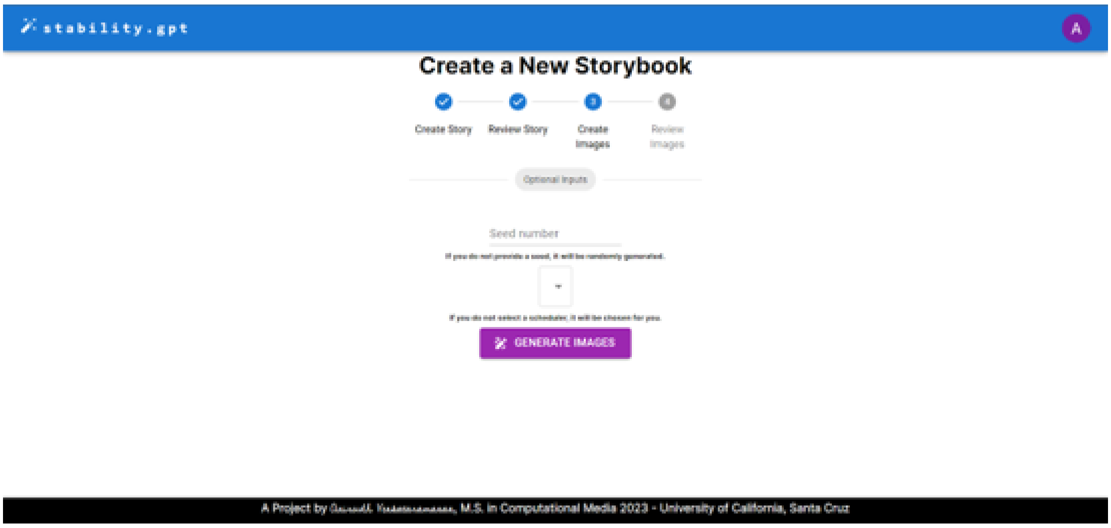
Figure 3.8: Panel where users can explain what changes they want made to the story that GPT-4 generated.

Users can choose to make changes to the story as many times as they like. It should be noted that the GPT-4 API is not free and more changes will result in more expenses. GPT-4 will remember the previous iterations of the story it generated and this is what makes updating the story possible. The prompt that allows the story to be changed is as follows, with the user input provided in parentheses - *Now, rewrite the exact same story in the same format as before and the exact same title (unless you are explicitly asked to change the title), but with the following changes - (changes provided by the user).*

While changes are being made, users will be shown the Loading screen from Figure 3.6. Once the user likes the story that the AI has generated, they can click the *Story Looks Good* button and move on to the next phase of creating a storybook - generating the images.

3.3 Generating consistent and thematically appropriate images with Stable Diffusion

The third step of the *Create a New Storybook* page is generating images with Stable Diffusion that have consistent characters and styles and that are thematically appropriate. Figure 3.9 shows the *Create Images* part of the *Create a New Storybook* page where users can optionally choose to provide a seed and scheduler for Stable Diffusion to use. If they don't provide this information, it will be randomly chosen for them. This information is optional because not everyone may be knowledgeable on what seeds and schedulers are. When I presented this project to a few students at my university, they were all confused on what seed number and scheduler to choose. Only those who have experience with schedulers and a variety of seed numbers will truly benefit from providing this information. A DDIM scheduler will have 12 inference steps as this is an optimal number of steps for high quality images [6]. All other schedulers will have the default 50 inference steps set by the Stable Diffusion API.



If you do not provide a seed, it will be randomly generated.

If you do not select a scheduler, it will be chosen for you.

A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.9: Third step of the *Create a New Storybook* page in stability.gpt is creating images. Users can enter a seed number and pick a scheduler to use if they wish. Then, they can click the *Generate Images* button.

Upon clicking the Generate Images button, users will once again be presented with the Loading screen from Figure 3.6. While the Loading screen is visible, GPT-4 first generates a Stable Diffusion prompt for each page of the story. The prompt that enables GPT-4 to do this in the correct format is as follows -

I want you to act as a Stable Diffusion Art Prompt Generator. The formula for a prompt is made of parts, the parts are indicated by brackets. The [Subject] is a living thing, place or inanimate object the image is focused on. [Emotions] is the emo-

tional look the subject or scene might have. [Verb] is What the subject is doing, such as standing, jumping, working and other varied that match the subject. [Adjectives] like beautiful, rendered, realistic, tiny, colorful and other varied that match the subject. The [Environment] in which the subject is in, [Lighting] of the scene like moody, ambient, sunny, foggy and others that match the Environment and compliment the subject. [Painting type] like oil painting, watercolor, pop-art, classicism, realism, photorealism and others. And [Quality] like High definition, 4K, 8K, 64K UHD, SDR and others. The subject and environment should match and have the most emphasis. It is ok to omit one of the other formula parts. For each page of the story, generate a prompt for a painting. Present each prompt as one full sentence, no line breaks, no delimiters, and keep it as concise as possible while still conveying a full scene from the page. If the [Subject] is a living being (human, animal, extraterrestrial life form), replace their name with the numerical age and name of a celebrity or fictional character they look like in all prompts they are present. You can also reference the movie or book a fictional character is from, but it must be done for all prompts where that character is present. If the [Subject] is an inanimate thing, reference a fictional or real-life object it looks like in all prompts where it is present. If the [Subject] is a place, reference a fictional or real life scene similar to it in all prompts where it is present. Assume that the reader of each prompt has no knowledge of previous prompts and avoid words like "same", "our", or "now" that reference other prompts. Also, if the [Subject] is a living thing, describe the clothes they are wearing right after you mention their name and before the [Emotions], including the color or design. The [Subject] should be wearing the same clothes *in all images unless the story text explicitly states they changed clothes or if a new outfit makes sense in the context of the story. The prompts should depict a scene from the story you just generated. They should all contain the exact same painting type. Determine the main [Subject] is for each page - either a person, place, or thing - and generate the prompt accordingly. Generally, if a person is holding an object, the [Subject] should be that object or thing. If a new area is being introduced for the first time, [Subject] should be a place. Have some variety. All of the prompts should not have a living thing as the [Subject]. Some prompts can have a location as the [Subject], showing a wider perspective of what is happening. Avoid words like "father", "son", "man", "woman", "mother", and "daughter" in your prompts. Instead, use the names of celebrities or fictional characters that look like them.*

Here is a sample output where the [Subject] is a person: "30 year old Rihanna, contemplative and reflective, sitting on a bench, cozy sweater, autumn park with colorful leaves, soft overcast light, oil painting painting style, 4K quality".

Here is a sample output where [Subject] is an object: "Titanic locket in the hands of 15 year old Tom Holland, shiny and bright, colorful, watercolor style, 4K quality".

Here is a sample output where [Subject] is a place - "Tweed Courthouse, bright and vibrant, with men and women in suits sitting in chairs, ambient light, photorealism style, 8K quality".

Create a json object with the prompts, with the property names being "PAGE #" where "#" is replaced by the page number. Make sure each prompt is one sentence long.

I made numerous changes to the above prompt to get it to its final form. I generated many Stable Diffusion prompts and images, updating the prompt based on things I did not like from GPT-4's output. My prompt initially did not ask GPT-4 to include the name of celebrities that the characters look like in its Stable Diffusion prompts. If the story used ambiguous terms like *boy* and *girl*, so would the Stable Diffusion prompts when referring to the characters. If a Stable Diffusion prompt referenced multiple characters with ambiguous terms, the generated image would often either generate one character with features of all the characters as shown in Figure 3.10 or mix up the features of each character as shown in Figure 3.11. It should be noted that the kandinsky scheduler was used for the images in Figures 3.10 and 3.11.

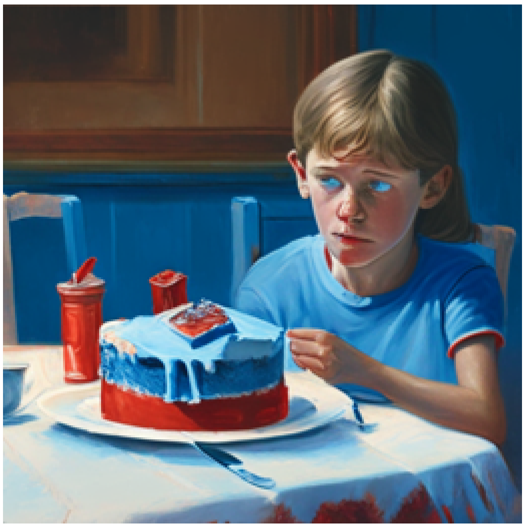


Figure 3.10: Stable Diffusion image for the prompt *A boy wearing a red shirt and blue jeans eating a cake at the dining table with his sister who is wearing a light blue dress.* Only one character is generated that has features of both the boy and the girl.

Given these results, I first decided to move away from using Stable Diffusion

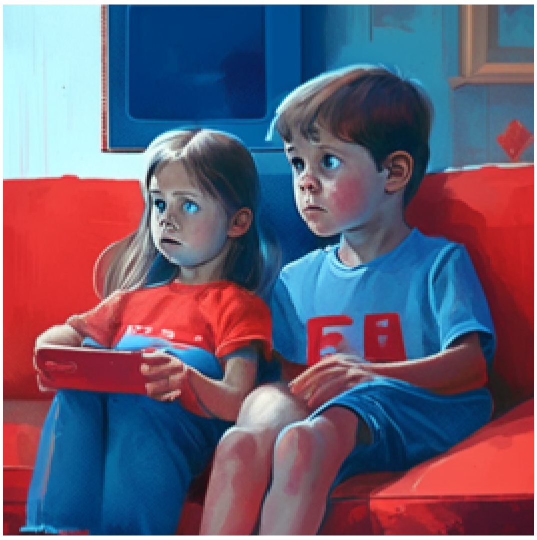


Figure 3.11: Stable Diffusion image for the prompt *A boy wearing a red shirt and blue jeans watching TV on the couch with his sister who is wearing a light blue dress.* While there are two characters generated, the colors of their outfits are swapped from the prompts with the boy wearing a light blue shirt and his sister wearing a red shirt. The sister is wearing her brother's outfit and not a dress.

completely and try using an AI generator that would create a video from a text prompt [3]. I had the idea of programmatically extracting frames from the video where the characters were the least deformed. However, I quickly realized that it would be difficult and time consuming to determine which frames to extract given the context of the story given that a video would contain tens, if not hundreds or thousands of frames.

Going back to Stable Diffusion, I found that including the names of celebrities in my prompts would result in the same characters being generated across all images [16]. There are thousands, if not millions of variations of boys, girls, men, women, fathers, and mothers that Stable Diffusion can generate, so including such a broad term can cause the model to get confused on what it should produce. On the other hand, there are only a few variations of someone like Hugh Jackman, so including a celebrity name will let Stable Diffusion know exactly what it needs to generate. I found that even when including the names of multiple celebrities in a prompt, Stable Diffusion would not get confused and generate each character correctly. This may be because the prompt had one character as the main subject while other characters were mentioned alongside the verbs describing what the main character was doing. I asked GPT-4 to replace the names of the characters in the story with the names of celebrities they looked like in the Stable Diffusion prompts and specified that it should avoid vague words at all costs.

A recommendation I received from a student at my university was to extend the approach of replacing a human character's name with a recognizable celebrity to animals, inanimate objects, and locations as well. Using the same logic as before, there are millions of ways to interpret words like book, drawer, house, park, and school. By referencing a location or object that exists in real life or a popular media and resembles what is in the story, I was able to get consistent locations and objects generated. This is why my prompt to GPT-4 specifies that all subjects - people, places, and objects - should not be referred to generically but rather as a real or fictional lookalike.

Another noteworthy aspect of my GPT-4 prompt is that it makes clear that words like *now*, *same*, and *our* that reference previous prompts or images should be avoided. This is because Stable Diffusion does not have any knowledge of what images it previously generated for a user. Stable Diffusion is trained on a large dataset of images from the internet and that is the only knowledge it has when generating images unless the model is fine-tuned [32][51]. Prompts may become repetitive by stating the same information again and again but that is only to the benefit of Stable Diffusion.

I noticed that an image would sometimes get flagged for having inappropriate content. By adding a negative prompt of *NSFW* (Not Safe for Work), I was able to get rid of this error. Negative prompts tell Stable Diffusion what it should not generate [56]. With a temperature of 0, GPT-4 was able to follow all of the requirements in my prompt and generate prompts in the right format, which I would then extract and put into Stable Diffusion to generate images.

I learned a lot about how Stable Diffusion prompts are generated from reading guides online and seeing prompts used by others to generate images of all sorts. It was through this and a post on an online forum about getting ChatGPT to generate Stable Diffusion prompts that I was able to talk about Stable Diffusion prompt structure in my GPT-4 prompt [5][60].

For each Stable Diffusion prompt, I generated four images, the maximum number allowed by the Stable Diffusion API [56]. Generative AI is not perfect, which is why I wanted users to choose what they liked best from a slew of options [63]. It would also help prevent users from having to spend more time and money in regenerating images. Figure 3.12 shows the benefit of generating four images from Stable Diffusion rather than one.

The prompt used to generate the images in Figure 3.12 is - *The International Space Station, bustling with activity, under the ambient light of various monitors and equipment, featuring a focused 20 year old Daniel Radcliffe conducting experiments in a space suit, oil painting style, 64K UHD quality.* Only the fourth image on the bottom right of Figure 3.3.4 shows Daniel Radcliffe's face in accordance with the prompt. Had only one image been generated, it could have potentially been one that is less consistent with the prompt. I would then have to spend more time and money trying to get the image I want. Providing more image options to users will help them make stories they like faster. It is more expensive to generate four images at a time as opposed to one, but the tradeoff is a much higher likelihood of improved image quality.

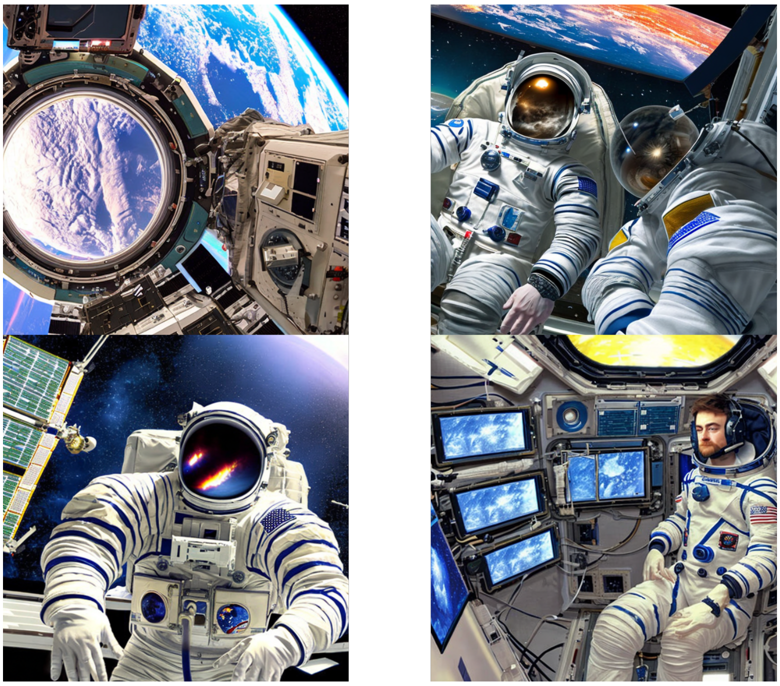
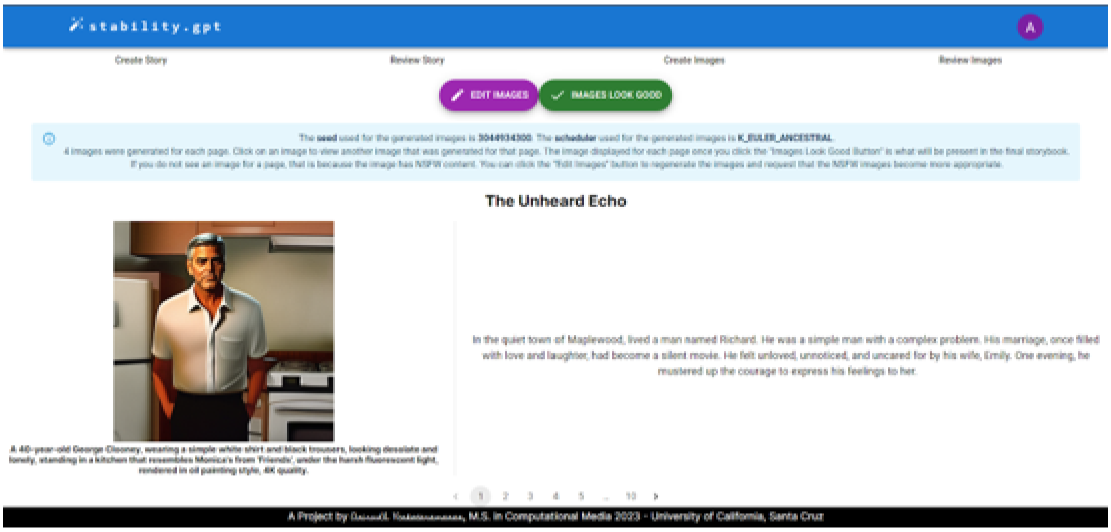


Figure 3.12: The four images generated by Stable Diffusion for the following prompt - *The International Space Station, bustling with activity, under the ambient light of various monitors and equipment, featuring a focused 20 year old Daniel Radcliffe conducting experiments in a space suit, oil painting style, 64K UHD quality.* Only the bottom right image shows Daniel Radcliffe's face as per the prompt.

A common criticism I received from students at my university who I presented stability.gpt to at this point was that all images were focused on a main character. Students mentioned that if a character was interacting with an object, they would have preferred the image to focus on that object rather than showing a closeup of the character. Similarly, the students wanted images to focus on locations more than living things and inanimate objects whenever a new location was first introduced. I made these changes by telling GPT-4 when the subject of the image should be a living thing, a lifeless object, and a location. I also added example prompts for when the subject was each of these. One student suggested I ask GPT-4 to reference objects and locations that exist in real life or in media (TV shows, movies, books) and that resemble the objects and locations in the story. This helped ensure that all parts of every image were consistently generated and not just the characters. With these updates, GPT-4 was able to generate prompts for each page where the focus was based on the storylines at that point in the narrative.

Once the images are generated, the Stable Diffusion API returns a temporary link from which they can be used [26]. I used this link to display the images alongside their prompts and the previously generated story text as shown in Figure 3.13. The seed and scheduler used to generate the images is also displayed in a light blue box below the two *Edit Images* and *Images Look Good* buttons. Users can click on an image to switch to another image that was generated for that same page. The image that is visible on the screen in each page is the image that will be in the final storybook.

The prompt for the image in Figure 3.13 for the first page of the book is - *A 40-year-old George Clooney, wearing a simple white shirt and black trousers, looking desolate and lonely, standing in a kitchen that resembles Monica's from 'Friends', under the harsh fluorescent light, rendered in oil painting style, 4K quality.* The prompt uses the name of a celebrity, George Clooney, instead of the name Richard in the story. It



The **seed** used for the generated images is **3044934300**. The **scheduler** used for the generated images is **K_EULER_ANCESTRAL**.
4 images were generated for each page. Click on an image to view another image that was generated for that page. The image displayed for each page once you click the "Images Look Good Button" is what will be present in the final storybook.
If you do not see an image for a page, that is because the image has NSFW content. You can click the "Edit Images" button to regenerate the images and request that the NSFW images become more appropriate.

The Unheard Echo

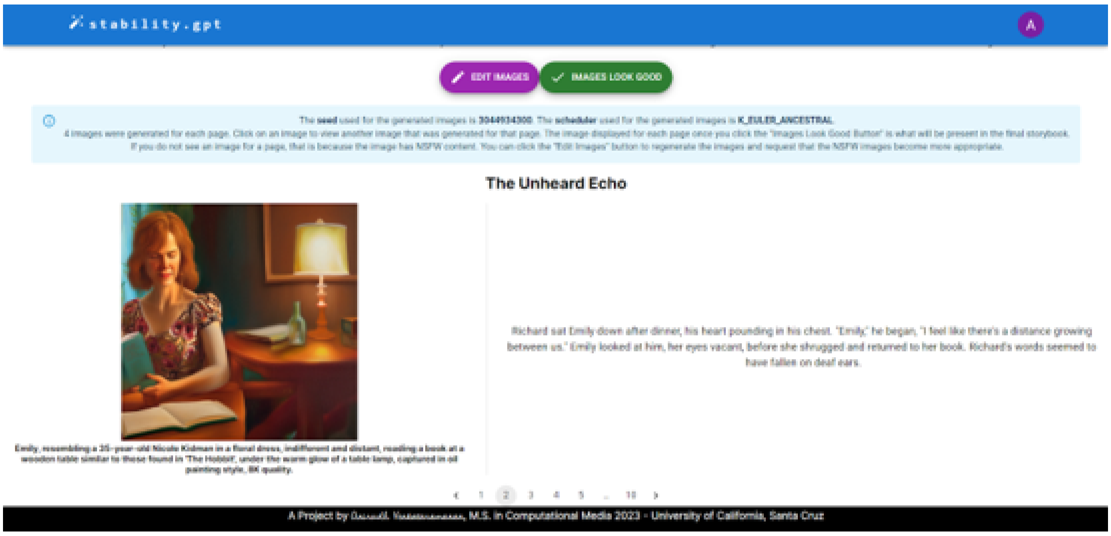
A 40-year-old George Clooney, wearing a simple white shirt and black trousers, looking desolate and lonely, standing in a kitchen that resembles Monica's from 'Friends', under the harsh fluorescent light, rendered in oil painting style, 4K quality.

In the quiet town of Maplewood, lived a man named Richard. He was a simple man with a complex problem. His marriage, once filled with love and laughter, had become a silent movie. He felt unloved, unnoticed, and uncared for by his wife, Emily. One evening, he mustered up the courage to express his feelings to her.

A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.13: Fourth and final step of creating a new storybook in stability.gpt - Reviewing the Stable Diffusion generated images. The images are displayed with their Stable Diffusion prompts and the story text. The seed and scheduler used to generate all the images are provided. Users can click on an image to view another image that was generated for that same page. The image that is on the screen is the image that will be in the final storybook.

also references a kitchen from a popular television show *Friends* so that Stable Diffusion knows exactly what kind of kitchen it should generate.



The **seed** used for the generated images is **3044934300**. The **scheduler** used for the generated images is **K_EULER_ANCESTRAL**.
4 images were generated for each page. Click on an image to view another image that was generated for that page. The image displayed for each page once you click the "Images Look Good Button" is what will be present in the final storybook.
If you do not see an image for a page, that is because the image has NSFW content. You can click the "Edit Images" button to regenerate the images and request that the NSFW images become more appropriate.

The Unheard Echo

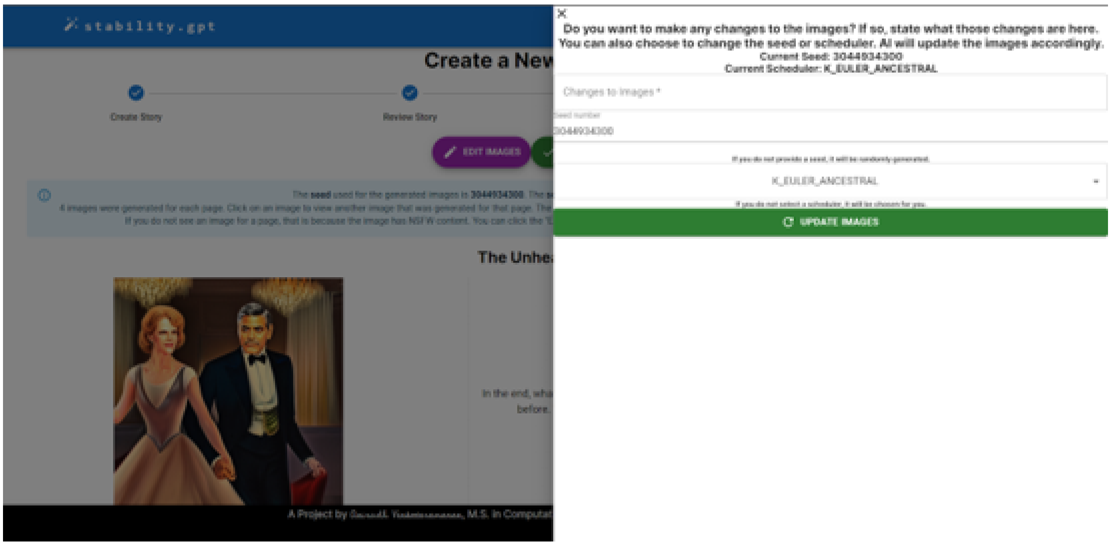
Emily, resembling a 35-year-old Nicole Kidman in a floral dress, indifferent and distant, reading a book at a wooden table similar to those found in 'The Hobbit', under the warm glow of a table lamp, captured in oil painting style, 8K quality.

Richard sat Emily down after dinner, his heart pounding in his chest. "Emily," he began, "I feel like there's a distance growing between us." Emily looked at him, her eyes vacant, before she shrugged and returned to her book. Richard's words seemed to have fallen on deaf ears.

A Project by [illegible], M.S. in Computational Media 2023 - University of California, Santa Cruz

Figure 3.14: Second page of the story from Figure 3.13.

The prompt for the image on second page of the story shown in Figure 3.14 is - *Emily, resembling a 35-year-old Nicole Kidman in a floral dress, indifferent and distant, reading a book at a wooden table similar to those found in 'The Hobbit', under the warm glow of a table lamp, captured in oil painting style, 8K quality.* Here, it is not a location, but rather an object - a wooden table - that is being compared to one in *The Hobbit* books and movies.



The seed used for the generated images is 3044934300. The s
4 images were generated for each page. Click on an image to view another image that was generated for that page. The
If you do not see an image for a page, that is because the image has NSFW content. You can click the 'E

In the end, wha
before.

A Project by [illegible], M.S. in Computat

Do you want to make any changes to the images? If so, state what those changes are here. You can also choose to change the seed or scheduler. AI will update the images accordingly.
Current Seed: 3044934300
Current Scheduler: K_EULER_ANCESTRAL

Changes to Images *

Seed number: 3044934300

If you do not provide a seed, it will be randomly generated.

K_EULER_ANCESTRAL

If you do not select a scheduler, it will be chosen for you.

Figure 3.15: Panel where users can make changes to the images that Stable Diffusion generated. The prompts, the seed, and/or the scheduler can be updated.

If a user wants to make changes to the generated images, they can do so in a window that slides in from the right of the screen, exactly like the window that comes up when editing the story text as previously shown in Figure 3.8. The window for updating images is shown in Figure 3.15. Users can choose to either edit some of the prompts, change the seed, choose a different scheduler, or do a combination of the three. As the images are being updated, the Loading page in Figure 3.6 will be displayed.

3.4 Publishing a Storybook

Once a user is satisfied with the book they generated with GPT-4 and Stable Diffusion, they can click the *Images Look Good* button. Upon doing so, the title, story text, images, image prompts, scheduler, seed number, and user email will be inserted into a SQL database.

With the storybook now stored in a database, users can access it from the dashboard and read it anytime.

3.5 Implementation Details

Stability.gpt was built using Next.js - a React.js framework - on the frontend, and Express.js - a Node.js framework - on the backend [1][37]. Both of these are web development frameworks I had prior experience in. Next.js also has a feature called the App Router, which makes creating new pages in a web application straightforward [11]. Since stability.gpt has multiple pages including a homepage, a dashboard, a page to actually create storybooks with AI, and a storybook viewer, Next.js made a lot of sense from a development standpoint.

I used the Planetscale database to store storybook data as it provides five gigabytes of storage for free [46]. The images are first stored in Cloudinary, a cloud-based image storage system so that they can be permanently accessible, and then their permanent public urls are posted to the database [14]. This is because the Stable Diffusion API only produces temporary links that expire after an hour [26].

Chapter 4

Evaluation

Stability.gpt was evaluated in the following three areas -

1. Updating the story text and images to the user's liking

2. Readability of story text

3. Consistency of style across all images

Stability.gpt's ability to update story text and images is crucial to ensuring the user is getting the storybook they want. If the user wants certain parts of a story or image changed, stability.gpt should be able to meet this criteria. Furthermore, if images are not consistent initially, stability.gpt should be able to fix that when it regenerates them (although ideally, the images are consistent from the get go).

The readability of the story and style consistency of images was measured for books in 10 different genres to see if there are certain types of storybooks stability.gpt is stronger at creating and that are more accessible.

4.1 Updating the Story and Images

To evaluate stability.gpt's ability to update the story text and images to satisfy users, I recruited five male individuals in their 20s to generate a story with stability.gpt. Once the story text was generated, four of them gave some feedback to the story, with the fifth being happy with how the generated story was. These changes were minor and involved changing names of a character or altering a small detail in the story as stability.gpt always followed the prompts set by each user. For three individuals, stability.gpt only had to update the story once before it pleased them. For the fourth individual, the story had to be regenerated twice. This individual was making a personal story about the downfall of one of their friends. Stability.gpt initially had a sentence that said the friend's downfall was live streamed. However, in real life, the downfall was not live streamed but rather just talked about on social media. When asked to change the sentence to reflect the real life scenario, the new sentence stability.gpt generated was - *Noah's downfall was not livestreamed but was the talk of social media.* Another change had to be requested, asking stability.gpt not to mention live streaming at all. To be fair, the update entered in the textbox that was sent to stability.gpt did say the updated sentence verbatim along with the page number where the change needed to take place. Had the update request specified that livestreaming should not be mentioned at all on that page, stability.gpt could have made all of the changes to this individual's desire in one update.

When it came to generating images, four of the five individuals were pleased with how the images generated initially turned out. As mentioned previously, stability.gpt generates four images per page so the user can select which one they like best. There were certainly some janky, out-of-place images generated for each page but ultimately there was at least one image generated for each page that had a consistent style with the other images. One individual who had generated a story about a cat chasing a bug had a critique that while the images generated followed the story's narrative, they weren't the highest quality. Upon regenerating the images, the individual lamented that they weren't coherent and that the cat and bug looked vastly different in each subsequent page. I also noticed that the only the prompt of the first image that features a character would include the name of the media that contains a lookalike (Puss in Boots from the *Shrek* movie for the cat and Flik from *A Bug's Life* for the bug). All subsequent prompts just had the replica's name without any mention of its source, making it difficult for Stable Diffusion to know what the character looks like. The individual did mention that the human character in the story - the cat's owner - was generated consistently in all the images she was in. It was just the animals that weren't up to their standards.

4.2 Story Text Evaluation

While the five individuals who created their own story liked the story text, it should be noted that they are all adults in their 20s with a more sophisticated reading comprehension level. Furthermore, not everyone who uses stability.gpt to generate stories will be generating stories for themselves; teachers or parents could be generating stories for their children, for instance. To evaluate the strength of stability.gpt's story text for a wider audience, I first asked ChatGPT to generate prompts for the ten most-selling book genres - Romance, Young Adult, Horror, Memoirs and Autobiographies, Science Fiction, Self-Help, Children's Literature, Fantasy, Crime and Mystery, and Historical Fiction [54]. I then asked stability.gpt to generate stories with these prompts, with each story being 10 pages and each title being generated by stability.gpt. I copied the generated story text to Readable, an online tool that calculates the readability of literary work to get feedback on how stability.gpt's story text could be improved to appease more individuals [50]. Table 4.1 contains the prompts generated by ChatGPT for each of the ten aforementioned genres as well as information on how accessible the stories generated from stability.gpt are. The ten stories can be read in full in the appendix.

Table 4.1: Story Prompts, Book Titles, and Flesch Reading Ease Score for 10 Genres

Genre	Prompt	Book Title	Flesch Reading Ease Score
Romance	Write about a bookshop owner discovering hidden love letters in an old tome.	The Bookstore's Secret Love Letters	63.2
Young Adult	In a world with unique powers, follow a teen discovering a forbidden ability.	The Forbidden Element	62.1
Horror	A family's pet dog acts strangely in their eerie new home.	The Shadows of the Past	62.1
Memoirs and Autobiographies	Share a world-famous musician's rise to stardom.	The Symphony of a Star	48.5
Science Fiction	A scientist teleports to an alien world.	The Accidental Voyager	49.2
Self-Help	An executive's chance meeting sparks a journey to inner peace.	The Executive's Journey	50.5
Children's Literature	Follow a brave rabbit rescuing friends from a wicked sorceress.	A Tale of Courage and Friendship	62.3
Fantasy	Narrate the story of a misfit dragon seeking legendary status.	The Misfit Dragon's Quest	60.5
Crime and Mystery	Solve a 1920s masquerade murder with an eccentric detective.	The Masquerade Murmur	70.7
Historical Fiction	A gladiator risks all for love and freedom in ancient Rome.	A Tale of Love and Freedom	56.5

Among many metrics, Readable most notably calculates the Flesch Reading Ease Score for any given piece of text. This score ranges from 0 - 100 with a higher score meaning that the text is easier to read. A score between 60-70 means that the text can be understood by the average young teenager. The score is calculated with a formula that considers the average length of sentences and average syllables in words. [19][61].

According to Table 4.2.1, the story from the Crime and Mystery genre had the highest reading ease score while the Memoirs and Autobiographies genre had the lowest reading ease score. The average for all ten stories was 58.56. This means that the stories are fairly difficult to read [61]. The biggest issue identified by Readable for all ten stories was that the sentences are too long, and this contributed towards the lower average Flesch Reading Ease Score. The students who tested stability.gpt in section 4.1 were all either college students or university graduates, making it easier for them to understand and approve the generated stories. The reading ease score shows that younger audiences may have a harder time comprehending stories generated from the system.

4.3 Image Generation Evaluation

Once the text had been generated for each story, stability.gpt generated images for each page. I used the DDIM scheduler with 10 inference steps as they lead to good quality images [6]. Of the four images generated for each page, I selected the ones I liked best. I did have to request a few image regenerations to try and maximize consistency. For instance, in the Crimes and Mystery book, a detective first meets a woman in a peacock outfit and then runs into her later on. Initially, the woman was only wearing the peacock outfit in the first image she was in. I had to request that the subsequent image with this woman also showed her in that same outfit. My request was honored and ChatGPT generated a new prompt which Stable Diffusion then used to generate an updated image. The woman now looks much more similar in both the images she is in.

I created a collage of the images for each story and conducted a survey where 12 individuals responded with whether the images for each story had a consistent style; if not, they explained why. The survey was sent out to students at my university and also posted on a forum for artificial intelligence enthusiasts. Figure 4.1 shows the gender composition of survey respondents. It should be noted that the average age of those who completed the survey is 32.4.

Figure 4.2 shows the results of the survey. The results show that images for the Science Fiction genre unanimously had consistency across all images. For both the Self-Help and Crimes and Mystery categories, only one respondent was in denial of consistent image generation. The reasoning provided was that the images with close-up characters were more realistic as opposed to images with a more wide angle which seemed more artificial.

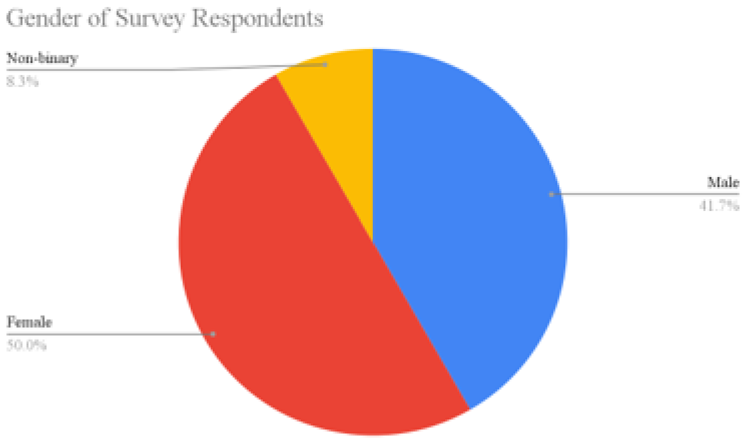


Figure 4.1: Pie Chart showcasing the gender distribution of the image consistency survey respondents.

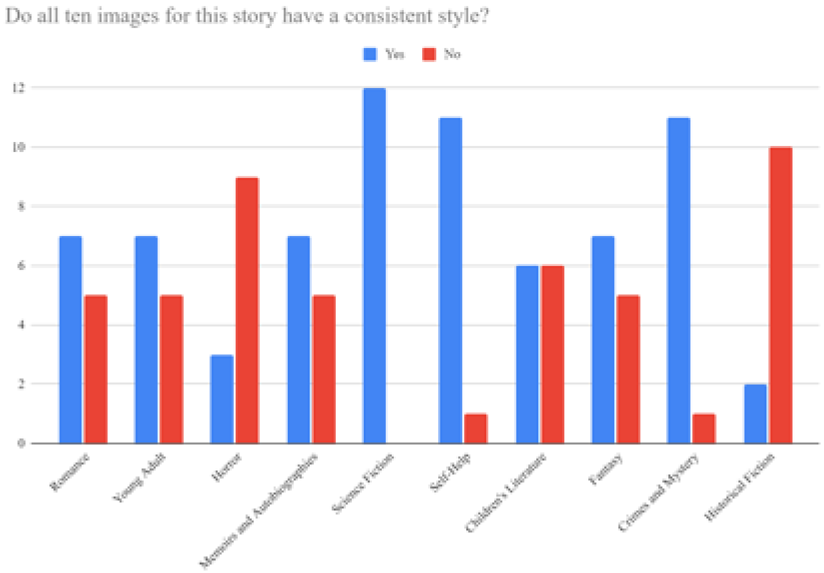


Figure 4.2: Results of the survey where participants looked at the images generated for 10 different stories - each of a different genre - and responded with whether or not they had a consistent style.

Moving towards the genres where the responses were more evenly divided. A common criticism for the romance story was that the tone/mood of the images was not consistent. The story does have some tonal shifts so there is a possibility that those who viewed the images along with the story text that previously felt the image styles were inconsistent would feel differently. For the Young Adult genre, the comments from naysayers varied, with some commenting on at least one image having a different style while others commented on groups of people (not any individual characters) looking different in one image versus the others. The children's story had six responses saying the images were generated consistently while the other six were in opposition. The most noteworthy comment was that the third image stood out from the rest. Looking back at the four images that were generated for this page, I admit I did choose an image that did not match stylistically with the images I chose for the other pages and that there were better options. The fantasy story also had mixed responses but ultimately only the final image was called out for being out of place. The story is about a tame dragon and his quest to be accepted by others. The final page talks about the dragon's story inspiring many others and shows a human being telling the story of the dragon to other creatures. This is vastly different from the other images where the dragon was one of the big focuses. It would have been more appropriate for the final image to show the dragon looking brave and proud.

There were two genres where the vast majority of responses were in favor of the images not having a consistent style. With the horror story, some said the dog in the story looked different in both the images it was in and that some images had a horror theme while others were more cheerful. While the story does have a positive ending and the last image is deservedly more vibrant, I agree that the other images could have had a more dreary look. For instance, the prompt for the first image referred to the Addams family house which is a horror house from a fictional horror franchise. However, the other prompts only generically referred to this location as a house without describing what it was like. The story that had the most objections to image style consistency was the historical fiction one. The biggest issue was that most images depicted a different time period. From the images alone, some thought there were multiple stories being visually presented and that the styles were also inconsistent.

Overall, while there were no comments provided on lack of consistency with human characters, the survey results show that the overall tone and style of images can be improved for most genres, as can the consistent generation of non-human characters. Currently, stability.gpt seems to be the best at generating science fiction storybooks.

Chapter 5

Conclusion

5.1 Discussion

It is evident that artificial intelligence and specifically LLMs and image generation models have the capability to generate creative content from a user prompt. Current methods to combine the power of both text and image generation models to push forward visual storytelling have resulted in inconsistent character generation and a disconnect between the story text and its paired image. This paper introduces a new system - stability.gpt - with the goal of generating storybooks that have consistent images that enrich the story text in an aesthetically pleasing way. GPT-4 generates (and updates) story text. Once it is to the user's liking, it generates Stable Diffusion image prompts for each page of the story which are then plugged into Stable Diffusion to generate the images. The story text and images are presented to users in a web browser in the frontend of stability.gpt. The user interface allows users to navigate between pages with a pagination bar and provide feedback to update the story and images.

The GPT-4 prompt that requests Stable Diffusion prompts clearly defines all the components that a Stable Diffusion prompt should have and keywords that are prohibited. In an attempt to maximize image consistency, the GPT-4 prompt asks that the Stable Diffusion prompts reference real and fictional locations, human beings, and objects - things that have an objective look and style. Each Stable Diffusion prompt starts by describing a subject - either a person, place, or inanimate object. The prompt sent to GPT-4 includes examples of prompts with all three subjects and explains how the subject for an image should be decided. This leads to a variety of angles and perspectives as opposed to a close-up of a character or object in all images.

By evaluating the system's ability to create storybooks of various genres, it is evident that stability.gpt's strengths currently lie in generating stories in the Science Fiction, Self-help, and Crimes and Mystery genres - genres that are typically more focused on one human character primarily. Characters are shown consistently in all images they are in as their celebrity lookalike is being generated. References to fictional or real world objects and locations also help images meet a certain aesthetic, although these references are not always present.

5.2 Future Work

Section 4.2 makes it clear that the sentences in the stories generated by stability.gpt are too long, making the story difficult for younger readers to comprehend. I definitely plan on adding an option for users to mention the intended audience of their book and asking stability.gpt to generate the story for that audience. For elementary school children, this would mean smaller sentences and simpler vocabulary.

As discussed in Section 4.3, the Stable Diffusion prompts generated by GPT-4 can still contain ambiguity. For example, if a prompt for an earlier page of the book refers to a house as a house from a popular movie or TV show, it will later go back to referring to that house as just a house with no mention of what the house should look like. This is especially apparent when the house is no longer the primary subject of the prompt. Furthermore, there are a few cases where the Stable Diffusion prompt will make the same reference to a subject from the real world or media in multiple prompts but the generated images will still have subjects that look different. I will work on updating my GPT-4 prompt to fix this issue with Stable Diffusion image prompt generation. I will specify that ambiguity should not be used under any circumstances and that references will always have to be made, regardless of if something is the subject of the image or a background prop. For animals and objects, I will also specify that the prompts must include physical descriptions such as color, size, and emotions, in addition to citing doppelgangers.

As mentioned in Chapter 1, artificial intelligence is thriving and is expected to grow massively in the coming years. As I was evaluating stability.gpt, a new and improved Stable Diffusion model called SDXL that generates even higher quality images was released. An API for SDXL is now available and I am excited to see if SDXL can further improve image quality and if so, by how much [45][56][57]. I have also recently gained access to the official Pathways Language Model (PaLM), the LLM that Google's Bard uses. One advantage of using this model instead of GPT-4 is that stability.gpt will be able to generate stories on current events as it is connected to the internet [15]. As new smarter LLMs and image generation models get released, I will be sure to integrate them into stability.gpt and see how they impact story generation. I am looking forward to seeing what the future of generative AI holds and how it impacts the ways stories are told.

Supplemental Files

The following supplemental files are available -

- PDF files of the ten books generated by stability.gpt for the evaluation of the system. There will be one PDF file per book and each PDF will include screenshots of every page for each story.
- The source code for stability.gpt along with instructions on how to run it locally on one's computer

Bibliography

[1] "About Node.Js." Node.Js, Node.js, nodejs.org/en/about. Accessed 1 July 2023.

[2] "The All-in-One Workspace for Your Notes, Tasks, Wikis, and Databases." Notion, Argil AI, argilai.notion.site/Introduction-to-automations-480be5fd297d4b0ca5c8cbe017d40e77. Accessed 3 Sept. 2023.

[3] "Andreasjansson/Stable-Diffusion-Animation." Replicate, replicate.com/andreasjansson/stable-diffusion-animation. Accessed 30 June 2023.

[4] Andrew. "ChatGPT: How to Generate Prompts for Stable Diffusion." Stable Diffusion Art, 25 Jan. 2023, stable-diffusion-art.com/chatgpt-prompt/.

[5] Andrew. "Stable Diffusion Prompt: A Definitive Guide." Stable Diffusion Art, 17 Aug. 2023, stable-diffusion-art.com/prompt-guide/.

[6] Andrew. "Stable Diffusion Samplers: A Comprehensive Guide." Stable Diffusion Art, 10 June 2023, stable-diffusion-art.com/samplers/.

[7] Ankita. "Stable Diffusion NSFW and Its Alternatives." MLYearning, 5 Aug. 2023, www.mlyearning.org/stable-diffusion-nsfw/.

[8] “Argil Live on the ChatGPT Plugin Store: Revolutionizing No-Code AI Automation.” Argil AI, www.argil.ai/blog/argil-the-chatgpt-plugin-for-image-generation. Accessed 3 Sept. 2023.

[9] Berkowitz, David. “Temperature Check: A Guide to the Best CHATGPT Feature You’re (Probably) Not Using.” LinkedIn, 30 Apr. 2023, www.linkedin.com/pulse /temperature-check-guide-best-chatgpt-feature-youre-using-berkowitz/.

[10] Borji, Ali. ”Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2.” arXiv preprint arXiv:2210.00586 (2022).

[11] “Building Your Application: Routing.” Building Your Application: Routing — Next.Js, Vercel, nextjs.org/docs/app/building-your-application/routing. Accessed 1 July 2023.

[12] Chanda, Prakriti. “Unlock ChatGPT’s Full Potential: How to Bypass ChatGPT Filter.” AMBCrypto, 4 Sept. 2023, ambcrypto.com/blog/unlock-chatgpts-full-potential-how-to-bypass-its-content-filter/.

[13] “ChatGPT Images Simplified: Argil Plugin for Seamless Generation.” Argil AI, www.argil.ai/blog/how-to-generate-chatgpt-images-using-argil-plugin. Accessed 3 Sept. 2023.

[14] Cloudinary, Cloudinary, cloudinary.com/. Accessed 1 Aug. 2023.

[15] Conway, Adam. “Google Bard: What Is It, and How Does It Work?” Developers, 27 Aug. 2023, www.xda-developers.com/google-bard/.

[16] Dance, Josh. “How to Create Consistent Characters in Stable Diffusion.” Mythical AI, Substack, 10 Feb. 2023, mythicalai.substack.com/p/how-to-create-consistent-characters.

[17] “Diffusion Models: A Practical Guide.” ScaleAI, scale.com/guides/diffusion-models-guide. Accessed 1 Sept. 2023.

[18] Duarte, Fabio. “Number of ChatGPT Users (2023).” Exploding Topics, Exploding Topics, 13 July 2023, explodingtopics.com/blog/chatgpt-users.

[19] “Flesch Reading Ease and the Flesch Kincaid Grade Level.” Readable, 9 July 2021, readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/.

[20] “Generative AI Market to Worth $63.05 Billion by 2028: Generative AI to Leave Biggest Impact on Drug Discovery, Material Science, and Financial Services.” GlobeNewswire News Room, SkyQuest Technology Consulting Pvt. Ltd., 9 Dec. 2022, www.globenewswire.com/en/news-release/2022/12/09/2571196/0/en/Generative-AI-Market-to-Worth-63-05-Billion-by-2028-Generative-AI-to-Leave-Biggest-Impact-on-Drug-Discovery-Material-Science-and-Financial-Services.html.

[21] “Gmail.” Information Technology Services, University of California, Santa Cruz, its.ucsc.edu/google/gmail.html. Accessed 1 Sept. 2023.

[22] “GPT-4.” OpenAI, 14 Mar. 2023, openai.com/research/gpt-4.

[23] Greenfield, Ely. “Adobe Firefly Generative AI Adobe Express to Inspire Millions to Create in Partnership with Google.” Adobe Blog, Adobe, 10 May 2023, blog.a

dobe.com/en/publish/2023/05/10/adobe-firefly-adobe-express-google-bard.

[24] Guinness, Harry. “Stable Diffusion vs. DALL-E 2: Which Is Better? [2023].” Zapier, Zapier, 5 May 2023, zapier.com/blog/stable-diffusion-vs-dalle/.

[25] “How to Write a Children’s Picture Book.” How to Write a Children’s Picture Book, Penguin, 1 June 2021, www.penguin.co.uk/articles/company-article/how-to-write-a-children-s-picture-book.

[26] “HTTP API Reference.” Replicate, Replicate, replicate.com/docs/reference/http#predictions.create. Accessed 4 July 2023.

[27] “Improving Storytelling through Human lt;gt; AI Collaboration. — Storybird.” StroyBord.Ai, storybird.ai/. Accessed 31 July 2023.

[28] Inglewood, Alex. “Guide: Stable Diffusion’s CFG Scale Explained.” Once Upon an Algorithm, 18 Mar. 2023, onceuponanalgorithm.org/guide-stable-diffusions-cfg-scale-explained/.

[29] Knight, Steven. “Most Popular Email Providers by Number of Users (2023).” SellCell, 15 Feb. 2023, www.sellcell.com/blog/most-popular-email-provider-by-number-of-users/.

[30] Kothari, Sneha. “Dall-E vs. Midjourney vs Stable Difussion: Which Is Better?” Simplilearn, Simplilearn, 23 June 2023, www.simplilearn.com/dalle-vs-midjourney-vs-stable-difussion-article?tag=DALL-E+vs.+Midjourney+vs+Stable+difussion%3A+Which+is+Better.

[31] "LLM Settings: Learn Prompting: Your Guide to Communicating with AI." Learn Prompting: Your Guide to Communicating with AI RSS, learnprompting.org/docs/basics/configuration_hyperparameters. Accessed 1 Sept. 2023.

[32] Laukkonen, Jeremy. "What Is Stable Diffusion? A Look at How One Artificial Intelligence Model Is Reshaping the Images You See." Lifewire, Lifewire, 9 May 2023, www.lifewire.com/what-is-stable-diffusion-7485593.

[33] "Libguides: Eng - Multimodal Texts: Mulitmodal - What Is It?" LibGuide, Danebank, 18 Oct. 2022, libguides.danebank.nsw.edu.au/c.php?g=874618&p=6279827.

[34] Meta Research. "Model Details." GitHub, 2023, github.com/facebookresearch/llama/blob/main/MODEL_CARD.md.

[35] "The Most Complete Guide to Stable Diffusion Parameters." OpenArt Blog, 13 Feb. 2023, blog.openart.ai/2023/02/13/the-most-complete-guide-to-stable-diffusion-parameters/.

[36] Nagy, Miklos, and Harish Prabhala. "The ML Developers Guide to Schedulers in Stable Diffusion." Segmind, 24 June 2023, blog.segmind.com/what-are-schedulers-in-stable-diffusion/.

[37] Next.Js by Vercel - The React Framework, Vercel, nextjs.org/. Accessed 1 July 2023.

[38] Nichols, Eric, Leo Gao, and Randy Gomez. ”Collaborative storytelling with large-scale neural language models.” Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games. 2020.

[39] Nori, Harsha, et al. ”Capabilities of gpt-4 on medical challenge problems.” arXiv preprint arXiv:2303.13375 (2023).

[40] “OpenAI Platform.” OpenAI Platform, OpenAI, platform.openai.com/docs/api -reference/introduction. Accessed 15 July 2023.

[41] “OpenAI Platform.” OpenAI Platform, OpenAI, platform.openai.com/docs/guides/gpt-best-practices/strategy-split-complex-tasks-into-simpler-subtasks. Accessed 15 July 2023.

[42] “OpenAI Platform.” OpenAI Platform, OpenAI, platform.openai.com/docs/guides/gpt/how-should-i-set-the-temperature-parameter. Accessed 1 Sept. 2023.

[43] Ouyang, Shuyin, et al. ”LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation.” arXiv preprint arXiv:2308.02828 (2023).

[44] Pavlik, John V. ”Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education.” Journalism Mass Communication Educator 78.1 (2023): 84-93.

[45] Podell, Dustin, et al. ”SDXL: improving latent diffusion models for high-resolution image synthesis.” arXiv preprint arXiv:2307.01952 (2023).

[46] “Pricing.” PlanetScale, planetscale.com/pricing. Accessed 10 July 2023.

[47] “Prompt Engineering: Learn Prompting: Your Guide to Communicating with AI.” Learn Prompting: Your Guide to Communicating with AI RSS, learnprompting.org/docs/basics/prompt_engineering. Accessed 15 June 2023.

[48] Raj, Gowtham. “What Is CFG Scale in Stable Diffusion and How to Use It.” DC, Decentralized Creator, 27 Feb. 2023, decentralizedcreator.com/cfg-scale-in-stable-diffusion-and-how-to-use-it/.

[49] Rangwala, Arva. “How to Use Midjourney API?” Open AI Master, 11 July 2023, openaimaster.com/how-to-use-midjourney-api/.

[50] Readable, Readable, 19 Apr. 2020, readable.com/.

[51] Schuhmann, Christoph, et al. ”Laion-5b: An open large-scale dataset for training next generation image-text models.” Advances in Neural Information Processing Systems 35 (2022): 25278-25294.

[52] Shakeri, Hanieh, Carman Neustaedter, and Steve DiPaola. ”Saga: Collaborative storytelling with gpt-3.” Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing. 2021.

[53] Short, Cole E., and Jeremy C. Short. ”The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation.” Journal of Business Venturing Insights 19 (2023): e00388.

[54] Singh, Soham. “Top 10 Most Selling Book Genres.” Mart, 6 Apr. 2023, gobookmart.com/top-10-most-selling-book-genres/.

[55] “Stability-Ai/SDXL.” Replicate, replicate.com/stability-ai/sdxl/api. Accessed 20 Sept. 2023.

[56] “Stability-Ai/Stable-Diffusion API.” Replicate, replicate.com/stability-ai/stable-diffusion/api. Accessed 15 July 2023.

[57] Stable Diffusion XL, stablediffusionxl.com/. Accessed 20 Sept. 2023.

[58] “Text to Cartoons — Storybird.” StoryBird.Ai, storybird.ai/text-to-cartoons. Accessed 1 Aug. 2023.

[59] “Usage Policies.” OpenAI, 23 Mar. 2023, openai.com/policies/usage-policies.

[60] “Using ChatGPT as a Prompt Generator -W/Example.” r/StableDiffusion, Reddit, 2 Mar. 2023, www.reddit.com/r/StableDiffusion/comments/11g9zul/using_chatgpt_as_a_prompt_generator_wexample/.

[61] van de Rakt, Marieke. “The Flesch Reading Ease Score: Why and How to Use It.” Yoast, 20 May 2019, yoast.com/flesch-reading-ease-score/.

[62] Wang, Zijie J., et al. ”Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models.” arXiv preprint arXiv:2210.14896 (2022).

[63] Wayner, Peter. “7 Dark Secrets of Generative Ai.” CIO, 12 Sept. 2023, www.cio.com/article/651570/7-dark-secrets-of-generative-ai.html.

[64] Witteveen, Sam, and Martin Andrews. ”Investigating prompt engineering in diffusion models.” arXiv preprint arXiv:2211.15462 (2022).