International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

StoryForge: AI-Powered Narrative with Dynamic Imagery and Voice

Dr. M. Thirunavukkarasu
Department of Computer Science Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Enathur, Kanchipuram, Tamil Nadu, India.
email address or ORCID

S R Bathrinathan
Department of Computer Science Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Enathur, Kanchipuram, Tamil Nadu, India.
11209A001@kanchiuniv.ac.in

Naga Skanda Kumar
Department of Computer Science Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Enathur, Kanchipuram, Tamil Nadu, India.
11209A009@kanchiuniv.ac.in

Yokesh M
Department of Computer Science Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Enathur, Kanchipuram, Tamil Nadu, India.
11209A024@kanchiuniv.ac.in

***Abstract*—StoryForge is an AI-powered narrative generation system that utilizes a combination of OpenAI's GPT-3.5 Turbo model [1], Hugging Face's Diffusion-based text-to-image generative model Stable Diffusion XL (SDXL) [4], and Hugging Face's Text-to-Speech Bark model [10] to create dynamic and engaging stories tailored to user-specified constraints. Users can input parameters such as genre, number of characters, character details, key events, mood, tone, point of view, and reader age to guide the narrative generation process. The system leverages the power of GPT-3.5 Turbo to generate the story text [1], SDXL to generate images for key events [4], and Bark to synthesize speech for the narrative [10], providing a comprehensive and immersive storytelling experience. A Streamlit-based user interface enables seamless interaction with the system, allowing users to input constraints, view generated text and images, and listen to the synthesized audio. It transforms user inputs into a multisensory narrative experience, marking a significant stride in the intersection of artificial intelligence and imaginative storytelling. StoryForge demonstrates the potential of AI to revolutionize storytelling, empowering users to create personalized and captivating narratives with ease and embark on a journey where words, images, and voices converge to create truly one-of-a-kind tales.**

***Index Terms*—StoryForge, GPT-3.5 Turbo, Bark, Stable Diffusion XL, Streamlit, Hugging Face, Generative Model, Storytelling, Multisensory**

I. Introduction

The primary objective of the StoryForge project is to develop an AI-powered narrative generation system that revolutionizes the storytelling experience. By seamlessly integrating OpenAI's GPT-3.5 Turbo model [1], Hugging Face's SDXL for text-to-image generation [4], and Bark for text-to-speech synthesis [10], the system aims to provide users with a comprehensive tool for creating dynamic and engaging stories. The project seeks to empower users by allowing them to input specific constraints, such as genre, character details, key events, mood, and more, guiding the AI in crafting personalized narratives. The goal is to showcase the potential of AI in transforming traditional storytelling methods, offering a user-friendly interface through Streamlit for a seamless and interactive experience.

Additionally, the project aims to demonstrate the versatility of AI in creative endeavors, pushing the boundaries of narrative generation by combining advanced language understanding [1], image creation [3], and speech synthesis [5]. By catering to diverse user preferences and providing a platform for the synthesis of text, images, and audio, StoryForge seeks to open new possibilities in content creation and storytelling, showcasing the capabilities of AI to enhance and personalize the creative process [7][12].

A. Scope of the Project

The scope of the StoryForge project encompasses a comprehensive and user-centric approach to narrative generation, aiming to redefine how stories are crafted and experienced. Firstly, the project will delve into the development and optimization of the AI architecture, leveraging the capabilities of OpenAI's GPT-3.5 Turbo model for natural language understanding [1]. This involves fine-tuning the model to interpret user inputs accurately and generate contextually relevant and engaging story text. The integration of Hugging Face's SDXL for text-to-image generation [4] and Bark for text-to-speech synthesis [10] will further enrich the storytelling experience, providing users with a multi-sensory journey.

Secondly, the scope extends to the creation of a user-friendly interface using Streamlit, ensuring seamless interaction and feedback between the user and the AI system [2]. This involves designing an intuitive platform where users can easily input their storytelling constraints, visualize generated text and images, and listen to synthesized audio. User feedback will be integral to refining the interface for accessibility and enhancing the overall user experience. The interface will serve as a bridge between the advanced AI technologies powering StoryForge and the creative intent of the users.

Lastly, the project aims to explore the scalability and adaptability of StoryForge to various storytelling genres and styles [6]. This involves testing the system across different narrative contexts, from fantasy to mystery, and adapting the AI algorithms to cater to diverse user preferences. The scope encompasses not only the technical aspects but also the creative possibilities that StoryForge can unlock, demonstrating its versatility in catering to a wide array of storytelling scenarios. Through this, StoryForge aspires to become a go-to tool for storytellers seeking a dynamic and personalized approach to narrative creation.

B. Existing System and Its Drawbacks

Limited narrative flexibility: The linear structure of traditional storytelling can restrict the exploration of alternative plotlines, character interactions, and narrative outcomes. This can hinder the creation of stories that are more intricate, unpredictable, and engaging for readers [11].

Reduced interactivity: The linear nature of traditional storytelling limits the potential for interactive storytelling experiences. Users are often confined to a passive role, unable to influence the story's progression or make choices that shape the narrative [2!].

Hindered exploration of multiple perspectives: Traditional storytelling often focuses on a single protagonist or a limited group of characters. This can limit the exploration of diverse perspectives, cultural backgrounds, and viewpoints, which can enrich the storytelling experience .

Challenges in incorporating nonlinear storylines: The linear structure of traditional storytelling can make it difficult to incorporate nonlinear plotlines, flashbacks, or parallel narratives. This can limit the ability to create stories that are more complex, layered, and thought-provoking [1].

Reduced adaptability to user preferences: Traditional storytelling often relies on predetermined plotlines and character arcs, which may not align with the specific preferences of individual readers. This can lead to stories that are less engaging and less relevant to the audience .

Limited immersion and sensory engagement: Traditional storytelling primarily relies on verbal or written descriptions to convey the narrative, which may limit the immersion and sensory engagement for the audience. Without visual or interactive elements, readers may struggle to fully visualize scenes or empathize with characters, potentially reducing the overall impact of the story experience.

II. Proposed Method

- Begin by conducting a thorough analysis of user requirements and expectations. Identify key parameters such as genre preferences, character details, mood, tone, and any other constraints users may want to specify.
- Integrate OpenAI's GPT-3.5 Turbo for natural language understanding and text generation. Fine-tune the model on a diverse dataset, emphasizing storytelling contexts to ensure it responds effectively to user prompts [1].
- Incorporate Hugging Face's SDXL for text-to-image generation [4]. Train the model to create relevant and visually appealing images corresponding to key events in the narrative.
- Integrate Hugging Face's Bark model for text-to-speech synthesis [5], allowing for the transformation of generated text into expressive and customizable speech.
- Develop a user-friendly interface using Streamlit to facilitate seamless interaction between users and the AI system [2]. Create an intuitive form where users can input their storytelling constraints easily.
- Implement real-time visualization features within the interface, enabling users to preview generated text and images, as well as listen to synthesized audio directly.
- Conduct rigorous testing to validate the system's performance across various scenarios, ensuring that it consistently generates coherent, engaging, and contextually relevant narratives [6].
- Establish a plan for continuous improvement, considering updates to underlying AI models, incorporating user feedback, and staying attuned to advancements in the field to enhance StoryForge's capabilities over time [9].

A. Architecture

As we start the Streamlit app, we land in our StoryForge's main page where the User has to enter the constraints, such as genre, character details, mood, tone, reader's age, key events, and Point-of-View. Once the constraints are submitted, they will be sent to GPT-3.5 Turbo model via API with the help of Langchain.

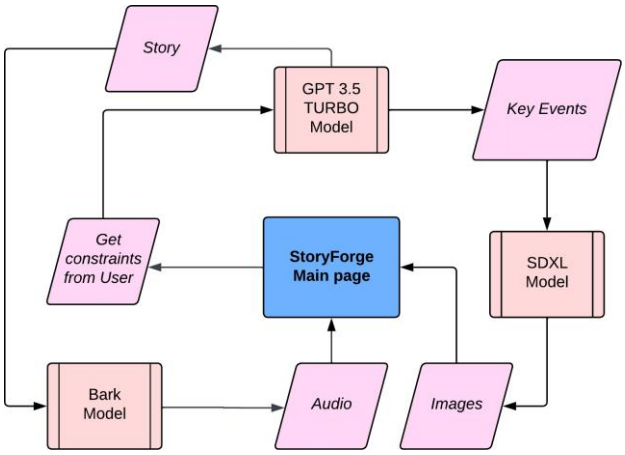

Fig. 1. StoryForge Architecture.

The GPT-3.5 Turbo model will generate a narrative, according to the prompt, that will fed to the Bark model for speech synthesis. The generated key events are sent to SDXL model for image generation.Finally, the story, image, and audio are displayed to the user in the Stramlit app for an immersive narrative experience.

Stable Diffusion XL model or SDXL is a Diffusion-based text-to-image generative model, which can be used to generate and modify images based on prompts passed as text.
It is a Latent Diffusion model, which uses two pretrained text encoders of which both are fixed. This has the ability

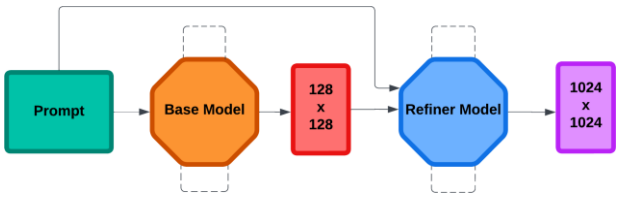

Fig. 2. SDXL Architecture.

to use Nvidia's CUDA toolkit, which enables us to use GPU-acceleration, and if our GPU VRAM is limited we have the option to enable CPU offload.

Bark is a text-to-speech is a model based on transformers. It can generate highly realistic, versatile vocalization in multiple languages, along with the ability to recreate various audio elements such as music, ambient sounds, and simple sound effects. The model can also produce nonverbal communications

Fig. 3. BARK Architecture.

like laughing, sighing and crying. It is a series of transformer models that work together to convert text to audio, namely, Text to Semantic Tokens, Semantic to Coarse Tokens, and Coarse to Fine Tokens with 10,000, 2x 1,024, and 6x 1,024 Output Vocab size respectfully.

III. ALGORITHM

1) **User Input:** Receive input such as constraints, including genre, character details, key events, mood, and tone, point of view, and target reader age from the user to guide the generation of narrative.

2) **Story Generation:** Utilize the GPT-3.5 Turbo API to generate a story by passing a prompt based on the constraints from the user to the API via LangChain's PromptTemplate and LLMChain. The template, defined as follows, is passed as an argument to PromptTemplate for the generation of the story:
template = """
you are a story teller;
you can generate a short story that is no more than 120 words;

CONTEXT: scenario
STORY:
"""

Algorithm 1 Algorithm for the Generate_Story function

```
Function GENERATE_STORY(scenario)
    template ← """
        you are a story teller;
        you can generate a short story that is no more than
    120 words;

        CONTEXT: scenario
        STORY:
    """

    prompt ← PromptTemplate(
        template=template,
        input_variables=["scenario"])

    story_llm ← LLMChain(
        llm=OpenAI(model_name="gpt-3.5-turbo",
        temperature=1), prompt=prompt, verbose=True)

    story ← story_llm.predict(scenario=scenario)
    return story
```

3) **Key Event Generation:** Again utilise the GPT-3.5 Turbo API but this time use the generated story as the prompt input and generate the list of Key events as output.

Algorithm 2 Algorithm for the Generate_ke function

```
Function GENERATE_KE(story)
    template ← """
        STORY: story
        give simple and small key events separated by ","
    """

    response ← openai.Completion.create(
        engine="gpt-3.5-turbo-1106",
        prompt=template,
        max_tokens=500,
        temperature=0.5,
    )

    generated_ke ← response['choices'][0]['text']
    return generated_ke
```

4) **Image Generation:** For each key event in the story generate an image using the SDXL model by passing the key events as prompts.
   a) **Latent Generation:** Use the base model to generate latents, usually accompanied by noise.
   b) **Denoising:** Use the refinement model to further process the latents, which are specialised for the final denoising steps.

Algorithm 3 Algorithm for the TXT2IMG function

```
Function TXT2IMG(prompt) {Convert text to image}

    pipe ← DiffusionPipeline.from_pretrained("stabilityai/
        stable-diffusion-xl-base-1.0",
        torch_dtype=torch.float16,
        use_safetensors=True, variant="fp16")

    pipe.to("cuda")
    images ←pipe(prompt=prompt).images[0]
    return images
```

5) **Audio Generation:** Use the Bark Model to synthesise speech for the generated narrative.
   a) **Semantic token generation:** The input text is tokenized with the BERT tokenizer from Hugging Face.
   b) **Coarse token generation:** The semantic tokens generated are used as input to give tokens from the first two codebooks of the EnCodec Codec.
   c) **Fine token generation:** The first two codebooks from EnCodec is fed to as input to present 8 codebooks from EnCodec.

Algorithm 4 Algorithm for the GEN_AUDIO function

```
Function GEN_AUDIO(story) {text to speech synthesis} 2:
processor ← AutoProcessor.from_pretrained("suno/ba 3:
rk")
    model ← AutoModel.from_pretrained("suno/bark")
    inputs ← processor( text=[story], return_tensors="pt")

    speech_values ← model.generate(**inputs,
                                   do_sample=True)

    sampling_rate ← model.config.sample_rate

    scipy.io.wavfile.write("bark_out.wav",
                           rate=sampling_rate,
            data=speech_values.cpu().numpy().squeeze())
```

6) **UI Showcase:** Use the Streamlit app to collective showcase all the generated text and media in an organised manner. Story, images, and audio are displayed using textarea, image, and audio functions, of the Streamlit library, respectively.

IV. Results

The Figure 4 showcases a user interface (UI) that features a form layout designed to collect various data inputs from the user for narrative generation. The form includes fields for key parameters such as the genre of the narrative, the number of characters, and detailed character information including their names, sexes, and ages.

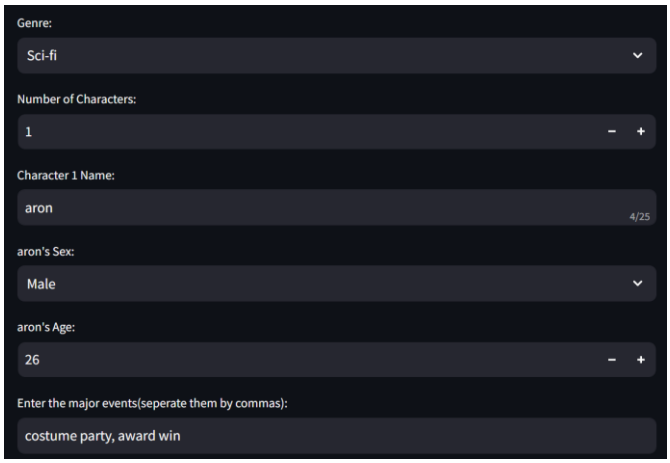

Fig. 4. constraints

Additionally, there is a section for the user to input major events of the narrative, with each event separated by commas. The UI is designed to be intuitive and user-friendly, facilitating the smooth input of data necessary for narrative creation.

Figure 5 extends the data collection process from Figure 4, incorporating additional parameters such as tone, mood, POV, and reader's age.

Fig. 5. constraints and generate story button

Upon completing the data entry, users can generate the story by clicking the "Generate Story" button, which triggers the generation and display of the narrative.

Figure 6 illustrates the generated story alongside a player interface enabling users to listen to the audio synthesized from the text.

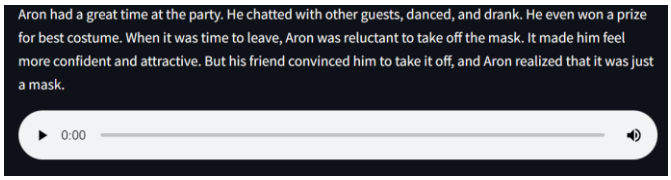

Fig. 6. generated story with synthesised audio and image

Fig. 7. generated image for key event ”party”

Fig. 9. generated image for key event ”win prize”

Fig. 8. generated image for key event ”feel more confident and attractive”

Fig. 10. generated image for key event ”friend convinced”

V. Conclusion

In conclusion, the StoryForge project represents a pioneering venture into the realm of AI-powered storytelling, aiming to provide users with a unique and immersive creative tool. By synergistically combining the language prowess of OpenAI’s GPT-3.5 Turbo [1], the visual storytelling capabilities of Hugging Face’s SDXL [4], and the auditory dimension brought by Bark’s text-to-speech synthesis [10], StoryForge aspires to redefine the way narratives are conceived and experienced. The user-friendly Streamlit interface acts as a bridge between advanced AI technologies and the creative intent of users, offering a seamless and interactive space for crafting personalized stories[14].

Through the development of StoryForge, we not only seek to showcase the potential of AI in transforming traditional storytelling methods but also to empower users to become storytellers in their own right. The adaptability of the system to diverse genres and styles, coupled with a user feedback loop for continuous improvement [8], positions StoryForge as a dynamic and evolving platform. As we delve into this

intersection of technology and creativity, StoryForge stands as a testament to the exciting possibilities AI brings to the world of narrative creation, fostering a new era where storytelling becomes a collaborative and personalized experience for users of all backgrounds and storytelling preferences.

REFERENCES

[1] M. Elgarf and C. Peters, "CreativeBot: a Creative Storyteller Agent Developed by Leveraging Pre-trained Language Models," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022, pp. 13438-13444, doi: 10.1109/IROS47612.2022.9981033

[2] M. Kamboj, C. Goyal, N. Ratra, "Unleashing potential: snowflake's streamlit strategy for genai solutions leveraging external network access and openai", Interantional Journal of Scientific Research in Engineering and Management, vol. 08, no. 03, p. 1-5, 2024. https://doi.org/10.55041/ijsrem29339

[3] R. Brisco, L. Hay, S. Dhami, "Exploring the role of text-to-image ai in concept generation", Proceedings of the Design Society, vol. 3, p. 1835-1844, 2023. https://doi.org/10.1017/pds.2023.184

[4] S. alam.A, N. Jeyamurugan, M. B, R. Veerasundari., "Stable diffusion text-image generation", Interantional Journal of Scientific Research in Engineering and Management, vol. 07, no. 02, 2023. https://doi.org/10.55041/ijsrem17744

[5] S. Joshi and V. Bairagi, "Recent trends in text to speech synthesis of indian languages", Helix, vol. 9, no. 3, p. 4931-4936, 2019. https://doi.org/10.29042/2019-4931-4936

[6] C. Wang, F. Pastore, A. Go¨knil, L. Briand, "Automatic generation of acceptance test cases from use case specifications: an nlp-based approach", IEEE Transactions on Software Engineering, vol. 48, no. 2, p. 585-616, 2022. https://doi.org/10.1109/tse.2020.2998503

[7] P. Atkinson and D. Barker, "Ai and the social construction of creativity", Convergence: The International Journal of Research Into New Media Technologies, vol. 29, no. 4, p. 1054-1069, 2023. https://doi.org/10.1177/13548565231187730

[8] D. Honeycutt, M. Nourani, E. Ragan, "Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy", Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, p. 63-72, 2020. https://doi.org/10.1609/hcomp.v8i1.7464

[9] J. Stray, "The ai learns to lie to please you: preventing biased feedback loops in machine-assisted intelligence analysis", Analytics, vol. 2, no. 2, p. 350-358, 2023. https://doi.org/10.3390/analytics2020020

[10] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhaoet al., "Multispeech: multi-speaker text to speech with transformer",, 2020. https://doi.org/10.48550/arxiv.2006.04664

[11] M. Riedl and R. Young, "From linear story generation to branching story graphs", Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, vol. 1, no. 1, p. 111-116, 2021. https://doi.org/10.1609/aiide.v1i1.18725

[12] T. Larsson, J. Font, A. Alvarez, "Towards ai as a creative colleague in game level design", Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, vol. 18, no. 1, p. 137-145, 2022. https://doi.org/10.1609/aiide.v18i1.21957