ISSN: 2473-4901
v11 n6429

An AI-based tool for Creating Social Stories for Children: NiteStory.AI

Asia Stevenson
stevensonasia@cityu.edu
MS in Computer Science (MSCS)

Sam Chung
chungsam@cityu.edu

School of Technology & Computing (STC)
City University of Seattle (CityU)

Abstract

Early childhood is a pivotal period for cognitive and emotional development, demanding effective tools to support self-regulation, social skills, and emotional well-being. Social stories have long served as valuable resources in this domain, yet their manual creation is time-consuming and resource-intensive, limiting accessibility and personalization. This paper introduces NiteStory.AI, an AI-powered storytelling tool that leverages Retrieval-Augmented Generation (RAG) in conjunction with multiple large language models (LLMs) to produce dynamically personalized social stories. By leveraging child-specific data—such as developmental milestones, interests, and personal challenges—the application generates contextually relevant and emotionally resonant narratives. This approach addresses existing limitations in AI storytelling tools, notably their limited emotional depth, lack of cultural sensitivity, and insufficient personalization. Our evaluation demonstrates significant improvements in readability and emotional appropriateness through iterative refinements of prompts to AI models. Metrics such as Flesch Reading Ease, BARTScore, VADER sentiment analysis, and ROUGE scores for RAG evaluation indicate enhanced readability, coherence, emotional positivity, and effective content integration from uploaded materials. The NiteStory.AI app thus provides a scalable and innovative solution, significantly reducing the burden of manual story creation and empowering parents and educators to support children's unique developmental journeys more effectively.

Keywords: AI-Based Storytelling Tool, AI-Supported Story-Writing, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Interactive Storytelling, Personalized Social Stories

GitHub: https://github.com/AsiShvets/NiteStory.AI

ISSN: 2473-4901
v11 n6429

An AI-based tool for Creating Social Stories for Children: NiteStory.AI

Asia Stevenson and Sam Chung

1. INTRODUCTION

Early childhood is a crucial period for developing self-regulation, emotional control, and social skills. Social stories support this growth by providing simple, relatable narratives that help children navigate various situations. However, creating personalized and engaging social stories for each child's can be time-consuming and demanding for parents and educators.

An AI-based storytelling app addresses this challenge by harnessing large language models (LLMs) and the power of Retrieval-Augmented Generation (RAG) to produce highly personalized social stories dynamically. By integrating child-specific developmental goals, interests, and challenges, this app ensures stories are deeply resonant and compelling for learning. . Visuals and interactive elements further enhance engagement and learning.

As Fang et al. (2023) note, AI-based tools now assist character development and plot creation, while also improving grammar and coherence. They reduce barriers to story creation by offering suggestions, analyzing writing styles, and enabling multimedia integration. Furthermore, these tools foster creativity, logic, and collaboration.

However, current AI-based story generators often struggle with achieving nuanced emotional depth and fail to fully grasp complex human experiences (Fang et al., 2023). This shortcoming can lead to stories that feel disconnected from children's realities and needs. This project specifically tackles these limitations by focusing on the creation of social stories for children — an approach proven particularly beneficial for learners with developmental or social-emotional challenges. By integrating personalization and the RAG mechanism, the proposed app aims to bridge the gap between traditional social storytelling and the transformative potential of AI-enhanced solutions.

NiteStory.AI combines multiple LLMs, a RAG layer, and a visual storytelling component to create highly personalized social stories for children. Unlike typical AI story generators, NiteStory.AI it adapts to each child's age, interests, and developmental goals, incorporating user-uploaded images to deliver to deliver meaningful and interactive experiences for families and educators.

2. BACKGROUND

Children experience rapid cognitive and emotional development during their early years, learning through new experiences and interactions with their environment. Parents and caregivers play a crucial role in guiding children as they navigate these experiences, helping them develop self-regulation skills—the ability to manage thoughts, emotions, and behaviors in various social situations. These skills are essential for fostering independence, emotional intelligence, and positive social interactions.

One effective tool for teaching self-regulation and social understanding is the social story — a structured narrative that clarifies social expectations and appropriate behaviors in real-life contexts. By offering clear, concise explanations of common scenarios, social stories enable children to process new experiences and respond in socially acceptable ways (Vollmer, 2023).

Early childhood is a formative period for developing social, emotional, and cognitive skills. Social stories have been widely recognized as an effective tool for teaching children how to navigate social situations and manage emotions (Gray, 2010). These stories are particularly beneficial for children with developmental delays, autism spectrum disorder, or social skill deficits, as they provide clear and structured guidance (Barry & Burlew, 2004). However, the process of tailoring these stories to each child's specific needs is often labor-intensive, which risks many children remaining underserved.

Although digital resources for story creation have become more prevalent, they often lack both personalization and interactivity. Whittingham et al. (2013) emphasize that customized and engaging learning tools significantly improve outcomes for children, underscoring the demand for a more adaptive solution. An AI-powered application capable of dynamically generating personalized social stories and incorporating RAG to retrieve contextually relevant information could greatly expand access to meaningful, individualized support for parents, educators, and children alike.

AI is transforming storytelling by enhancing creativity, generating novel ideas, and accelerating the writing process. Writers can use AI-powered tools to brainstorm plotlines, create characters, and even produce visuals, making storytelling more dynamic and immersive. AI also speeds up drafting, allowing authors to focus on refining their narratives rather than starting from scratch. Additionally, AI contributes to other storytelling mediums like video games and films, adapting plots to user interactions and streamlining scriptwriting. By blending technology with human imagination, AI opens new possibilities for storytelling across various formats.

3. RELATED WORK

The integration of artificial intelligence (AI), especially large language models (LLMs), into storytelling applications has garnered significant attention in recent years. These technologies have the potential to revolutionize the creation, personalization, and application of stories across various contexts, such as education, mental health, and emotional development. This section summarizes findings from important research papers to examine the use of AI-powered storytelling, the strengths and limitations of current approaches, and the implications for future developments.

De Lima et al. (2023) introduced ChatGeppetto, an AI-powered storyteller designed to create interactive and co-creative storytelling experiences. The study emphasizes how ChatGeppetto utilizes large language models (LLMs) to enhance engagement through dynamic narratives. The advantages of this approach include its ability to adapt to user input and generate coherent, contextually rich stories. However, challenges remain, such as maintaining long-term narrative coherence and addressing ethical considerations in AI-generated content. This research is significant as it highlights the potential of LLMs to transform narrative creation in both entertainment and education.

Progga et al. (2024) investigated the use of large language models (LLMs) in personalized storytelling for postpartum wellbeing. Their study shows that AI-generated narratives can effectively address the emotional and psychological needs of postpartum individuals by creating supportive and relatable stories tailored to their experiences. A significant strength of this research is its emphasis on mental health applications, highlighting the therapeutic value of storytelling. However, the study also notes limitations in ensuring cultural sensitivity and inclusivity in AI-generated content. Overall, the findings enhance our understanding of how AI-driven storytelling can support mental health initiatives and promote emotional resilience.

Seo et al. (2024) introduced ChaCha, an AI application aimed at encouraging children to express their emotions through storytelling. By utilizing large language models (LLMs), ChaCha prompts children to reflect on personal experiences and articulate their feelings. The research highlights the effectiveness of co-creative storytelling in fostering emotional development and enhancing communication skills in children.

The strengths of this approach include its focus on child-centered design and its adaptability to individual needs. However, potential weaknesses involve the risk of over-reliance on AI, which may reduce opportunities for human interaction. Overall, this study provides valuable insights into the intersection of AI and emotional education.

Zhang et al. (2024) introduced Mathemyths, an AI-driven storytelling tool designed to teach mathematical language through collaborative storytelling between children and AI. The study highlights how co-creative narratives can enhance learning engagement and comprehension in mathematical contexts. The strengths of this approach include its innovative use of storytelling to simplify complex concepts and its potential to foster collaborative learning. However, there are challenges, such as ensuring the accuracy of mathematical representations and balancing creative and instructional elements. This research is significant for exploring the role of AI in education and its implications for collaboration between children and AI.

These studies collectively highlight the various applications of large language models (LLMs) in storytelling, spanning areas such as mental health, emotional development, education, and entertainment. They emphasize the transformative potential of AI-driven narratives while also addressing challenges like ethical considerations, cultural sensitivity, and the necessity for user-centered design. These references are essential to my study, as they establish a foundation for understanding the current landscape of AI storytelling and its future possibilities.

4. APPROACH

This AI-based storytelling app integrates multiple AI models and a RAG approach to create engaging, personalized stories for children. The app allows users to upload images and text inputs, which are processed to generate customized narratives.

Unlike traditional story generators that rely solely on text inputs, this solution leverages image-to-text processing and RAG for deeper personalization. The table below highlights key differences. Table 1 below shows key differences between existing story generators and this solution.

Feature	Existing Story Generators	This Approach
Input Type	Text-based only	Text and image-based, with optional retrieving the context from PDF
Personalization	Limited (user-supplied details)	Uses PDFs for context-aware stories
AI Models	Single model (e.g., GPT-3.5)	Multiple LLMs (GPT-3.5, GPT-2, BLIP, Ollama)
Local AI Model Support	Cloud-based only	Supports local processing with Ollama

Table 1: Comparison of my Approach with Existing Solutions

To ensure the AI-based storytelling app meets the needs of users, the following requirements are identified:

- Image Upload & Processing: Users should be able to upload images to generate stories based on visual input.
- Text-Based Story Generation: Users should provide text prompts for custom story generation.
- Personalized Stories: The system should integrate user-provided PDFs (e.g., favorite books, learning materials) to generate personalized stories.
- Offline & Online Functionality: Users should have access to cloud-based LLMs when available and a local AI model when offline.
- Secure & Scalable Storage: Image and text data should be securely stored and easily retrievable.
- Interactive & Responsive UI: The web interface should be intuitive and support real-time interactions.

System Design & Implementation

This approach enhances AI storytelling by incorporating image-based inputs and personalized content retrieval. By using multiple AI models and a hybrid online/offline processing capability, the system ensures flexibility, engagement, and scalability. The NiteStory.AI storytelling tool contains the following layers:

User Interaction Layer (Frontend)

The process begins with user unteraction through a React.js frontend. Parents and educators can upload images or PDFs and generate stories.

- Technology: React (JavaScript)
- Functions:
  - Upload images and PDFs
  - Input text prompts for story generation
  - Displays generated stories dynamically
  - Calls FastAPI backend for processing

Backend (FastAPI)

The FastAPI backend handles various API endpoints.

- Technology: Python (FastAPI framework)
- Functions:
  - Handles requests from the front end and routes them to appropriate AI models.
  - Provides REST API endpoints:
    - /api/upload-image: Converts uploaded images to text using a vision model
    - /api/upload-pdf: Processes PDF file as a context for RAG
    - /api/generate-story-from-image: Generates a story based on text prompts and enhances story generation with RAG
    - /api/evaluate-story: Evaluates quality of the story

AI Processing Layer

At the core of the system is the AI Processing Layer, which integrates multiple language models to enhance storytelling. For deeper personalization the RAG model extracts relevant content from user-uploaded PDFs.

- Hugging Face BLIP LLM: Converts images to text descriptions
- OpenAI GPT-3.5 LLM: Generates high-quality stories
- Hugging Face GPT-2 LLM: Alternative story generator with “Young-Children-Storyteller-Mistral-7B” model specifically trained for storytelling

- Ollama LLM: Provides local model processing when cloud models are unavailable
- Retrieval-Augmented Generation (RAG): Enhances personalization by integrating data from uploaded PDFs

Storage & Configuration
The system securely manages configuration through environment variables, ensuring sensitive API keys remain protected.

- Environmental Variables: .env files store API keys securely
- AWS S3: Stores uploaded images and PDFs for persistent access

Design diagram of this approach is pictured in Figure 1 of the Appendix. Figure 2 of the Appendix describes the sequence diagram – the flow of the requests and responses between the layers of the system.

Testing Strategy for the AI-Based Storytelling App
To ensure the reliability and user satisfaction of the AI-based storytelling app, we will implement a comprehensive testing strategy covering functional, non-functional, and AI-specific testing. This approach will help identify bugs, ensure seamless integration between components, and verify the accuracy of AI-generated content.

Functional Testing
Functional testing verifies that each feature of the app works as expected. By means of functional testing, we can ensure seamless data flow between the frontend, backend, and AI models. Test scenarios would include image uploads triggering proper story generation.

Usability Testing
To ensure the user interface (UI) is intuitive and user-friendly, we will gather feedback from real users through usability testing sessions. This process will involve observing how users interact with the app, identifying any areas where they encounter difficulties, and collecting their suggestions for improvement.

AI-Specific Testing
We will ensure the accuracy, relevance, and fairness of AI-generated stories through a structured testing process. First, we will validate the quality of stories produced by models like GPT-3.5, Young-Children-Storyteller-Mistral, and Ollama, ensuring they meet the desired standards. Additionally, we will test the effectiveness of Retrieval-Augmented Generation (RAG) by confirming that personalized stories accurately reflect the content from user-uploaded PDFs. This will include testing edge cases, such as PDFs with complex formats, to ensure the system can handle a variety of scenarios.

Example Test Cases
Table 2 below shows the example test cases that are expected to be performed.

Test Case	Input	Expected Result
Image Upload Validation	Upload an image	Image uploads successfully, triggers story generation
Invalid File Handling	Upload an executable file (.exe)	The system rejects the file with an error message
Story Generation Accuracy	1. Input text: "A dragon in space" 2. Upload an image of a truck.	1. The story includes a dragon and space-related elements 2. The story contains details about a truck.
API Error Handling	Simulate API downtime	A user receives a friendly error message
RAG Personaliza tion	Upload the child's PDF (favorite book)	The story includes personalized elements from the PDF

Table 2: Test Cases

This multi-layered testing strategy ensures the app is robust, secure, and delivers a high-quality storytelling experience.

5. DATA COLLECTION

In compliance with the requirement to avoid human-subject data collection (and thus forgo the need for Institutional Review Board approval), this project exclusively utilizes secondary sources of information and system-generated metrics. Specifically, any textual or image-based examples that inform or test the AI storytelling system are drawn from publicly available academic, governmental, or professional databases and websites (e.g., open-access image repositories and educational text corpora). These resources offer content representative of real-world scenarios without involving direct interaction with human subjects.

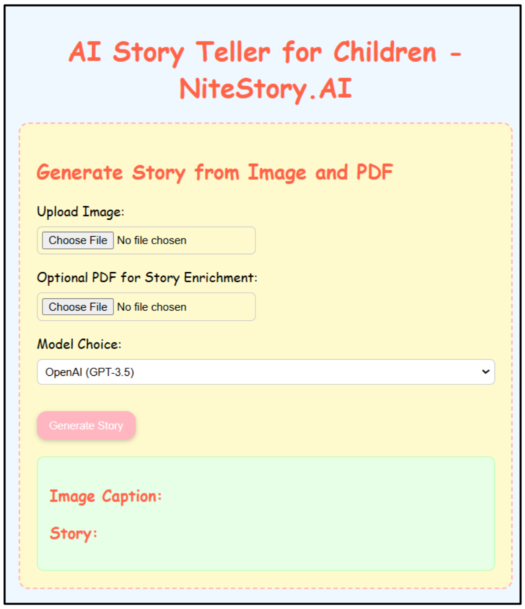
To evaluate the performance of the AI storyteller, system-generated metrics are collected during functional and AI-specific testing phases. Key performance indicators include accuracy (e.g., correctness in referencing the retrieved context) and coherence of narratives.

Input data

The tool has the following input data, which is shown in Figure 3:

1. Uploaded picture.
   - Purpose: Visual input that could help set context, mood, or theme for the story (e.g., a child's favorite character, an inspiring scene, etc.).
   - Usage: Although not strictly required, the picture can guide the tone or specific elements within the story.

2. (Optional) PDF Document
   - Purpose: Serves as a reference, containing a child's background information, plot elements, or specific vocabulary to be integrated into the story. Example: a favorite book.
   - Usage: The information extracted helps align the story with the provided material, ensuring consistency, depth, and high personalization.



AI Story Teller for Children - NiteStory.AI

Generate Story from Image and PDF

Upload Image:

Choose File No file chosen

Optional PDF for Story Enrichment:

Choose File No file chosen

Model Choice:

OpenAI (GPT-3.5)

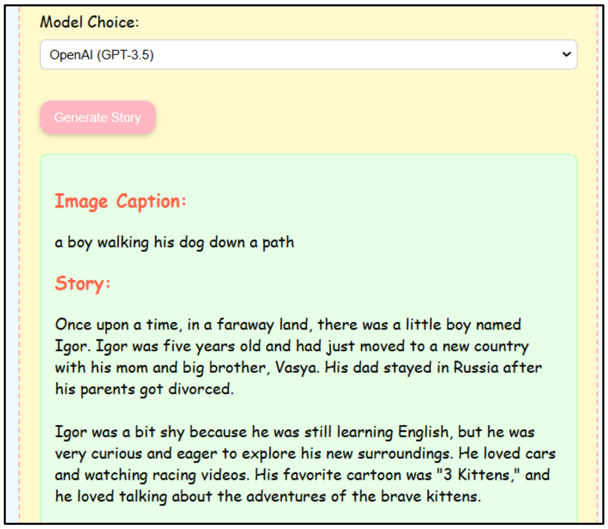
Generate Story

Image Caption:

Story:

Figure 3: Input Data on the UI

3. LLM (Large Language Model)
   - Purpose: The core engine for text generation, utilizing advanced language-processing capabilities to craft coherent, creative, and contextually relevant narratives.
   - Usage: Combines inputs from the picture, PDF (if any), and personalization parameters to generate the final, customized story output.



Model Choice:

OpenAI (GPT-3.5)

Generate Story

Image Caption:

a boy walking his dog down a path

Story:

Once upon a time, in a faraway land, there was a little boy named Igor. Igor was five years old and had just moved to a new country with his mom and big brother, Vasya. His dad stayed in Russia after his parents got divorced.

Igor was a bit shy because he was still learning English, but he was very curious and eager to explore his new surroundings. He loved cars and watching racing videos. His favorite cartoon was "3 Kittens," and he loved talking about the adventures of the brave kittens.

Figure 4: Output on the UI

Output data

Output data in Figure 4 is a personalized story (text and voice).

- Definition: A unique, context-aware narrative tailored to the child's profile and any relevant details from the PDF or picture.
- Characteristics: Engaging, age-appropriate language, thematically consistent with reference materials, and enriched by the LLM's ability to integrate all inputs into a cohesive story.

6. DATA ANALYSIS

The analysis of data collected from the AI-based storytelling app focuses on assessing the quality of generated stories and the effectiveness of personalization.

Personalization Accuracy

To evaluate the app's ability to personalize stories based on user-provided context, the generated narratives are compared with input parameters, including:

1. Contextual Alignment: How well the story incorporates elements from the uploaded image (e.g., elements from the image).
2. Character Consistency: Whether recurring characters or story elements appear consistently across multiple story generations.

3. RAG Effectiveness: Analysis of how effectively the Retrieval-Augmented Generation (RAG) module integrates relevant details from uploaded content (interests, favorite book themes).

Manual reviews are used to analyze the alignment between input prompts and generated stories. A relevance score is assigned based on how well the AI adheres to user-provided details.

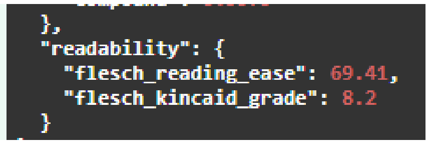
AI-Generated Narrative Quality

To assess the quality of AI-generated stories, both subjective and objective measures are used:

Readability Score is computed using standard readability indices (Flesch-Kincaid readability score) to ensure the story is appropriate for the child's age.

For children's books, a desirable Flesch-Kincaid readability score typically falls within the range of "80-100", which translates to a reading level suitable for early elementary schoolers, indicating a very easy-to-read text with simple sentence structures and vocabulary. (Posts, 2022).

Initially, Flesch Reading Ease was 69.41 which indicates the text is moderately easy to read, but for very young children (the tool's target audience), stories need a higher score (typically 80 or above) to ensure easier comprehension. Initial Flesch-Kincaid grade level was 8.2. This suggests that the text is written at about an 8th-grade level. For a children's story, especially one aimed at younger kids, one would ideally target a lower grade level to match their reading abilities. Figure 5 shows initial measurements.

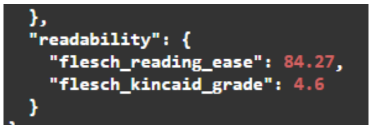


```
  },
  "readability": {
    "flesch_reading_ease": 69.41,
    "flesch_kincaid_grade": 8.2
  }
```

Figure 5: Initial Readability Score

So, the adjustment of the prompt instructions to guide the language model toward producing stories that are easier to read for younger children was necessary.

After testing the request to LLMs was modified to achieve appropriate level for little children (see Figure 6).



```
  },
  "readability": {
    "flesch_reading_ease": 84.27,
    "flesch_kincaid_grade": 4.6
  }
```

Figure 6: Improved Readability Score

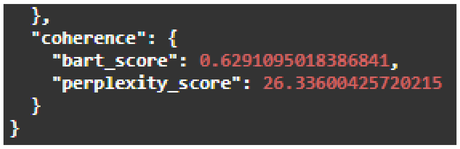
New readability scores indicate a strong improvement in making the story accessible:

- Flesch Reading Ease (84.27): A score in the low 80s means the text is very easy to read, which is excellent for young readers.
- Flesch-Kincaid Grade Level (4.6): A grade level of about 4.6 suggests that the content is appropriate for younger children around the 1-3rd grade, so this level is quite good for many children.

Overall, these metrics suggest that the story is engaging and easy to understand. These scores are a solid benchmark for readability in a children's storyteller app.

Coherence and Flow. Evaluated through natural language processing (NLP) techniques such as BARTScore and Perplexity Score that measure sentence structure, logical transitions, and overall narrative cohesion (see Figure 7).

- BARTScore predicts sentence-level coherence and fluency.
- Perplexity Score assesses the fluency of generated text by measuring how predictable the next word is within a sentence, with lower values indicating better coherence. (Papaoikonomou, 2024)



```
  },
  "coherence": {
    "bart_score": 0.6291095018386841,
    "perplexity_score": 26.33600425720215
  }
}
```

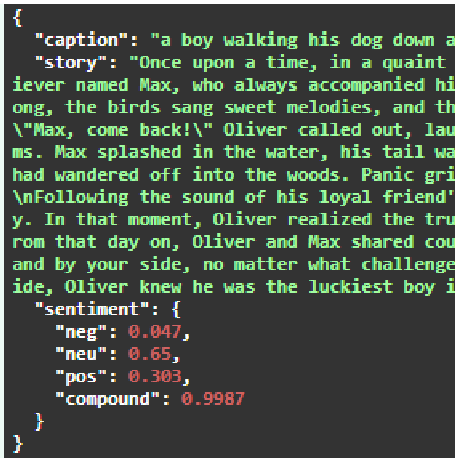
Figure 7: Coherence and Flow Metrics

BARTScore is a metric that leverages the BART sequence-to-sequence model to evaluate generated text by computing its reconstruction likelihood, effectively measuring its fluency and coherence. It stands for Bidirectional and Auto-Regressive Transformer (Yuan et al., 2021)

BARTScore of the story in the case study is ≈ 0.63. In this framework, a higher score suggests that the text is more coherent and fluent. A score of 0.63 indicates moderate coherence. It means that while the text has a reasonable flow and logical transitions, there's potential room for improvement when compared to texts with higher BARTScores.

Perplexity (often abbreviated as PPL) stands as one of the most common metrics for assessing Language Models (LLMs). Calculating perplexity necessitates having access to the probability distribution for each word generated by your model. It is a measure of how confidently the model is able to predict the sequence of words. The higher the perplexity, the less confidently the model predicts the observed sequence. (Papaoikonomou, 2024)

Perplexity Score in the case study is ≈ 26.33. Lower perplexity values imply that the text is more predictable and, therefore, typically more fluent. A perplexity of 26.33 shows that the text is moderately fluent.



```
{
  "caption": "a boy walking his dog down a
  "story": "Once upon a time, in a quaint
iever named Max, who always accompanied hi
ong, the birds sang sweet melodies, and th
\"Max, come back!\" Oliver called out, lau
ms. Max splashed in the water, his tail wa
had wandered off into the woods. Panic gri
\nFollowing the sound of his loyal friend'
y. In that moment, Oliver realized the tru
rom that day on, Oliver and Max shared cou
and by your side, no matter what challenge
ide, Oliver knew he was the luckiest boy i
  "sentiment": {
    "neg": 0.047,
    "neu": 0.65,
    "pos": 0.303,
    "compound": 0.9987
  }
}
```

Figure 8: Sentiment Analysis – VADER Technique

Emotional Depth. Assessed using sentiment analysis tools (VADER - Valence Aware Dictionary and sEntiment Reasoner) to determine whether the story conveys appropriate emotions, such as excitement, empathy, or reassurance. (GeeksforGeeks, 2024). See Figure 8.

For AI storyteller app aimed at children, these sentiment values serve as quality indicators to ensure the emotional tone is appropriate and engaging. In the above figure 4 this is how each metric can be interpreted:

- Negative (neg: 0.047): This indicates that only about 4.7% of the language in the story carries negative sentiment. For children's stories, you generally want this number to be low so that the narrative isn’t too harsh or upsetting.
- Neutral (neu: 0.65): About 65% of the text is neutral. Neutral sentiment often corresponds to descriptive or narrative parts of the story. This is normal as a story needs plenty of exposition to set the scene and provide context.
- Positive (pos: 0.303): Around 30.3% of the language is positive, reflecting cheerful or uplifting expressions. In a children’s story, having a healthy amount of positive sentiment helps create an engaging and friendly tone.
- Compound (compound: 0.9987): The compound score is a combined measure that ranges from -1 (very negative) to +1 (very positive). A score close to +1— as in this case —shows that the overall tone of the story is extremely positive. This is especially desirable in content for kids, as it ensures that the narrative is encouraging and reassuring.

These metrics set thresholds to flag stories that might have too much negative content or insufficient positive sentiment. For example, if a story’s compound score is significantly lower than expected, it might need adjustments to better suit a young audience.

While some conflict (a slight negative component) is necessary for an engaging narrative, the overall positive tone should prevail by the story’s end. These values help ensure that even if there’s a moment of tension, it is quickly resolved positively.

As the AI storyteller is refined in the future, these sentiment scores can be used to compare different versions of generated stories.

RAG Implementation Assessment

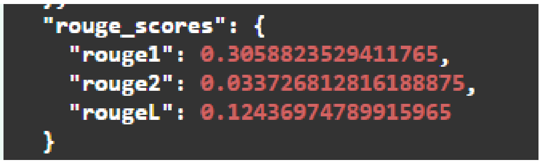
A common evaluation framework for Retrieval-Augmented Generation (RAG) systems is RAGAS. However, in this project, I chose not to use RAGAS for several key reasons.

RAGAS is typically used when there is a clear ground truth—an expected "correct" output that the generated content should align with.
In our case, the AI-generated stories are not meant to be direct summaries or fact-based answers, but rather creative narratives inspired by the uploaded PDFs. Since there is no single "correct" version of the story, RAGAS metrics like faithfulness and answer correctness are not applicable.

Unlike traditional RAG applications where retrieved content must be strictly factual and aligned with the source, NiteStory.AI is designed to generate engaging and imaginative stories based on themes, concepts, or characters from the PDF rather than copying its exact content.
Using RAGAS would unfairly penalize creative variations, even though those variations are intentional and beneficial for storytelling.

Instead, we chose another key metric for evaluating the effectiveness of the RAG implementation. And it is ROUGE (Recall-Oriented Understudy for Gisting Evaluation). While ROUGE is not directly part of the RAGAS framework, it is a well-established method for evaluating text similarity and summarization quality.

ROUGE in Figure 9 allows us to measure how much of the key content from the PDF is reflected in the generated story while still allowing for creative differences.
It helps us strike a balance—ensuring that the generated stories retain relevant elements from the PDFs without enforcing strict textual overlap. ROUGE measures the overlap of n-grams, word sequences, and word pairs between a reference text and a generated text. In this case, ROUGE is used to compare the retrieved PDF content (or its summary) with the generated story to determine if key elements from the PDF are present. (Papaoikonomou, 2024)



```
  "rouge_scores": {
    "rouge1": 0.3058823529411765,
    "rouge2": 0.033726812816188875,
    "rougeL": 0.12436974789915965
  }
```

Figure 9: ROUGE Score

For instance, the evaluation of our case study yielded the following results:

- ROUGE-1 score: 0.31, indicating that approximately 31% of the individual words (unigrams) in the generated story match those in the PDF content.
- ROUGE-2 score: 0.03, showing that only 3% of consecutive word pairs (bigrams) are shared between the two texts.
- ROUGE-L score: 0.12, reflecting that the longest common subsequence between the generated story and the PDF text captures roughly 12% overlap.

These results suggest that while the generated story incorporates a fair number of key words from the PDF (as evidenced by the ROUGE-1 score), it retains only limited matching phrases (ROUGE-2) and longer structural similarities (ROUGE-L). This indicates that while the model successfully integrates essential elements of the source content, it does not directly copy extensive phrases or narrative structures. Thus, the generated stories maintain originality while still being informed by the provided reference material.

The results from these analyses provide insights into personalization effectiveness and story quality. Based on these findings, iterative improvements are proposed to enhance the AI model's adaptability, storytelling depth, and user engagement.

This data-driven approach ensures that the AI-based storytelling app continues to evolve in delivering meaningful, personalized, and engaging narratives for children.

7. FINDINGS

The evaluation of NiteStory.AI demonstrated its effectiveness in generating personalized and engaging social stories for children by leveraging Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and image-to-text processing. Key findings from the study highlight the strengths, challenges, and areas for improvement in AI-generated storytelling.

The readability analysis showed that AI-generated stories initially had a Flesch Reading Ease score of 69.41 and a Flesch-Kincaid grade level of 8.2, indicating that the text was moderately complex and not well-suited for young children. However, by adjusting the prompts to emphasize simpler language, shorter sentences, and age-appropriate vocabulary, the readability improved to 84.27 (Flesch Reading Ease) with a Flesch-Kincaid grade level of 4.6, making the stories significantly more accessible to young readers.

To assess the fluency and logical structure of generated stories, BARTScore and Perplexity Score were used. These findings suggest that while stories are well-structured and understandable, improving transitions and narrative consistency could further enhance engagement.

Sentiment analysis using VADER confirmed that stories maintained a positive and emotionally engaging tone, which is crucial for content aimed at children. Compound sentiment score: 0.9987 – a high positive score, confirming the uplifting and reassuring nature of the stories.

These results indicate that NiteStory.AI successfully generates stories with an emotionally supportive tone, suitable for young audiences.
The ROUGE metric was used to assess how effectively the generated stories incorporated content from user-uploaded PDFs (e.g., favorite books, learning materials). The results showed that while the generated stories successfully integrate key concepts and themes from PDFs, they do not simply copy text verbatim. This aligns

with the goal of using RAG to personalize narratives while maintaining originality.

Using Hugging Face BLIP, images were successfully converted into text prompts, which were then used as the basis for story generation. This feature enhanced personalization by allowing children's favorite characters or settings to be included in the story. The evaluation confirmed that image-based storytelling significantly improved engagement and contextual relevance. These findings highlight the strengths and effectiveness of NiteStory.AI in generating personalized social stories. The system successfully integrates LLMs, RAG, and image-to-text models to produce engaging, emotionally supportive, and developmentally appropriate narratives for children. Areas for improvement include enhancing coherence and fluency, refining retrieval-based personalization, and exploring additional evaluation techniques to further optimize storytelling quality. These insights will guide future enhancements to improve AI-driven storytelling experiences.

8. CONCLUSION

This project aimed to simplify the creation of personalized social stories for children, focusing on emotional management, self-regulation, and social understanding. Traditional social stories, while effective, are time-intensive and lack real-time adaptability. Despite advances in AI-generated storytelling, many systems struggle with emotional nuance and child-specific personalization.

To address these challenges, we developed NiteStory.AI, an AI-powered storyteller integrating multiple Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and image-to-text processing. Our goals were:

- Personalization – Leveraging user-uploaded PDFs (e.g., books, therapy materials, personal story) and images to craft emotionally resonant stories.
- Practicality & Engagement – Ensuring accessible, uplifting narratives with appropriate reading levels.

NiteStory.AI effectively integrates PDFs and images to produce engaging, relevant stories. Metrics such as Flesch Reading Ease, VADER sentiment, and ROUGE for RAG validate the system's ability to generate meaningful content. The system tailors language complexity and sentiment for young audiences, ensuring a cheerful and emotionally safe experience.

The development of NiteStory.AI was constrained by time, limiting the ability to conduct extensive iterative fine-tuning of LLMs. As a result, while the system performs well in generating personalized narratives, further refinements could enhance fluency and coherence. Additionally, the project did not involve direct user testing; instead, story quality and emotional impact were evaluated using automated metrics and expert reviews rather than real-world child-parent interactions. This lack of live feedback leaves room for further validation and optimization.

NiteStory.AI demonstrates that AI-driven storytelling can successfully blend RAG, LLMs, and image-to-text processing to create personalized, emotionally supportive narratives for children. By addressing key challenges in social story creation, this project highlights the potential of AI while paving the way for future research and enhancements.

9. FUTURE WORK

While NiteStory.AI successfully demonstrates the potential of AI-driven storytelling, several areas can be further explored to enhance its effectiveness, personalization, and ethical considerations.

A key area for future improvement is security and content moderation. Ensuring that all generated images and textual content are appropriate for children is crucial. Implementing stricter content filtering mechanisms, leveraging AI-powered image moderation, and incorporating parental controls would help maintain a safe and age-appropriate experience.

Additional personalization features would further refine the user experience. Expanding customization options, such as allowing users to select a child's or character's gender, age, and specific interests, could make stories more relatable and engaging. Introducing real-time adaptive storytelling — where narratives dynamically adjust based on a child's mood, preferences, or previous interactions — could significantly enhance immersion and emotional connection.

From a technical perspective, optimizing offline functionality by incorporating smaller, fine-tuned LLMs for edge devices would improve accessibility, particularly for users in lowconnectivity areas. Further refinements in RAG retrieval strategies and prompt engineering could enhance coherence, fluency, and emotional depth in generated stories.

Additionally, conducting direct user testing with children and caregivers would provide valuable insights into engagement, comprehension, and emotional impact. A study involving real-world feedback would help refine storytelling approaches and validate the system's effectiveness in fostering social and emotional learning.

By addressing these areas, NiteStory.AI could evolve into a more personalized, safe, and adaptive storytelling tool that better supports children's emotional growth and learning.

10. REFERENCES

Barry, L. M., & Burlew, A. (2004). Social Stories: A Strategy for Social Skills Instruction for Children with Autism Spectrum Disorder. Journal of Special Education, 38(2), 111-118.

De Lima, E. S., Feijó, B., Cassanova, M. A., & Furtado, A. L. (2023). ChatGeppetto - an AI-powered Storyteller. In ACM International Conference Proceeding Series (pp. 28–37). https://doi.org/10.1145/3631085.3631302

Fang, X., Ng, D. T. K., Leung, J. K. L., & Chu, S. K. W. (2023). A systematic review of artificial intelligence technologies used for story writing. Education and Information Technologies, 28(11), 14361–14397. https://doi.org/10.1007/s10639-023-11741-5

GeeksforGeeks. (2024, December 11). Sentiment Analysis using VADER Using Python. GeeksforGeeks. https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/

Gray, C. (2000). The new social story book. https://ci.nii.ac.jp/ncid/BA53520228

Papaoikonomou, Q. H. T. V. T. (2024, October 9). RAG evaluation metrics: UniEval, BLEU, ROUGE & more - Elasticsearch Labs. Elasticsearch Labs. https://www.elastic.co/search-labs/blog/evaluating-rag-metrics

Posts, V. M. (2022, January 8). How to assess the readability of your book. shannonmeyerkort.com. https://shannonmeyerkort.com/2022/01/08/how-to-assess-the-readability-of-your-book/#:~:text=The%20Flesch%20Reading%20Score%20gives,of%20Wrath%20only%20scores%20680L

Progga, F. T., Khan, A., & Rubya, S. (2024). Large Language Models and Personalized Storytelling for Postpartum Wellbeing. CSCW Companion '24: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing, 653–657. https://doi.org/10.1145/3678884.3681921

Seo, W., Yang, C., & Kim, Y. (2024). ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 903. https://doi.org/10.1145/3613904.3642152

Vollmer, E. V. (2023, February 9). Using Social Stories to Improve Your Child's Understanding & Behavior | TherapyWorks. https://therapyworks.com/blog/language-development/home-tips/using-social-stories-improve-childs-development/

Whittingham, K., Rinehart, N., & O'Kearney, R. (2013). Personalized Learning Tools and Their Impact on Engagement and Learning Outcomes in Children with Autism Spectrum Disorders. Journal of Autism and Developmental Disorders, 43(2), 329-337.

Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating Generated text as Text Generation. Neural Information Processing Systems, 34. https://dblp.uni-trier.de/db/journals/corr/corr2106.html#abs-2106-11520

Zhang, C., Liu, X., Ziska, K., Jeon, S., Yu, C., & Xu, Y. (2024). Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling. CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 274. https://doi.org/10.1145/3613904.3642647

Appendix

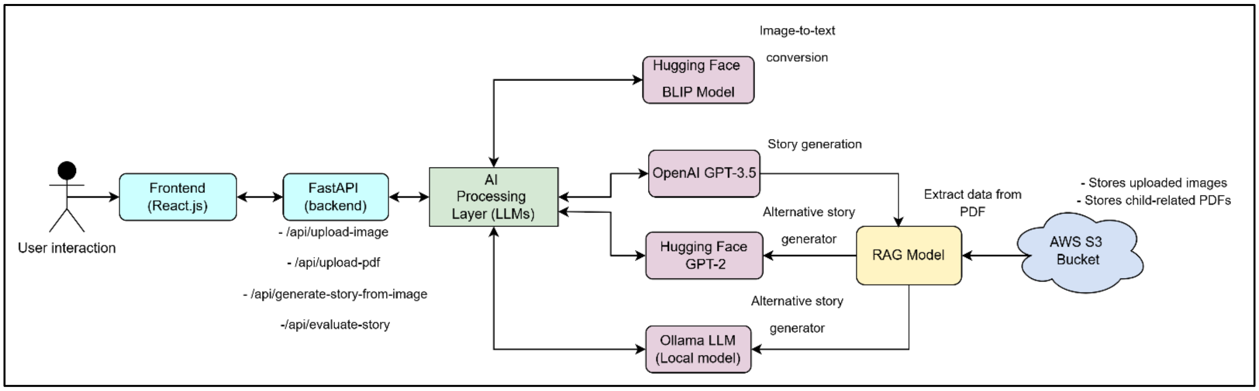


Figure 1: Design of the Approach

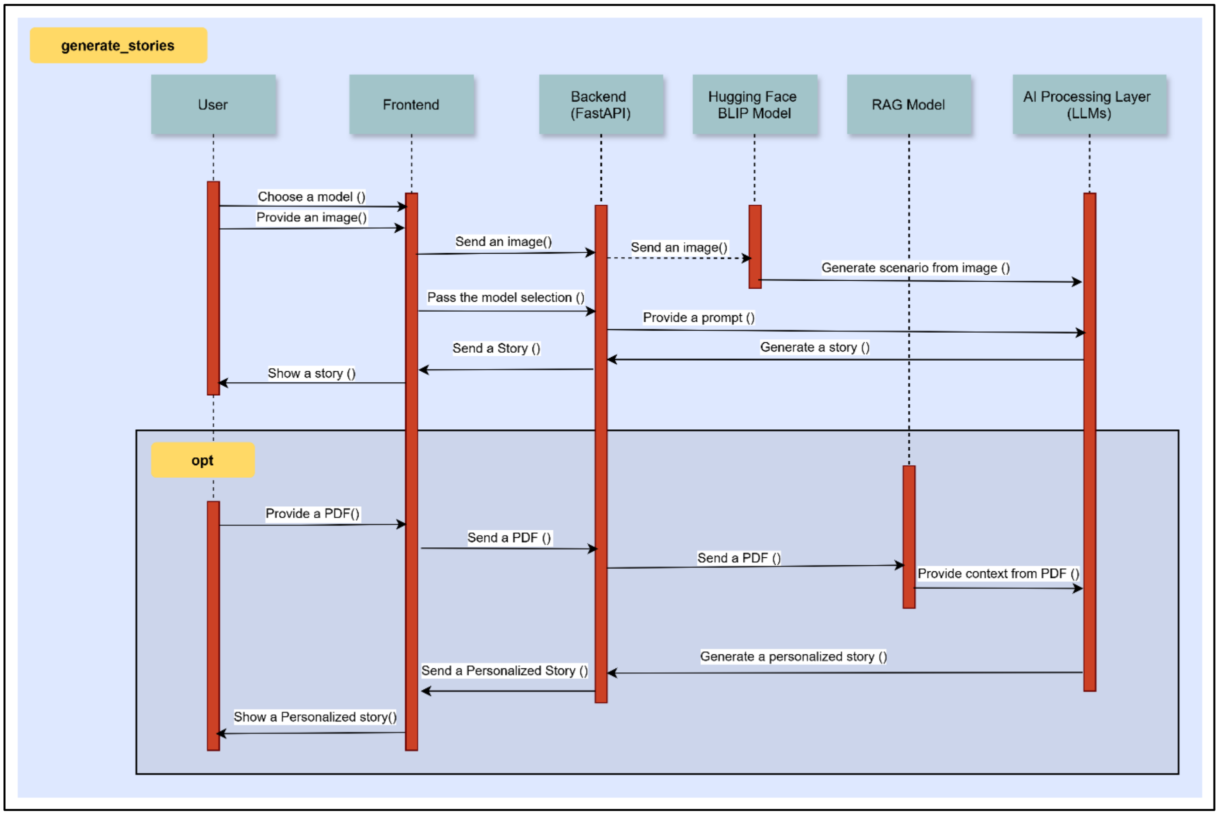


Figure 2: Sequence Diagram of the System