Anushka Satav, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# AI-Powered Personalized Storybook Generator – Using Multi-Modal Content Generation

**Anushka Satav, Siddhi Shinde, Ashish Singh, Supriya Jagtap**
Dept. of Information Technology

**Abstract- This paper showcases an AI-driven system that creates and shares beautifully illustrated kids' storybooks via a seamless, all-in-one multimodal workflow. Kick off with a basic user prompt and chapter count, and a large language model (LLM) delivers a ready-to-go story: a snappy title, cover idea, and a neat lineup of chapters each packed with engaging text and tailored illustration prompts. By crafting the whole multi-chapter arc in one shot, it locks in smooth, consistent storytelling from start to finish. Spot-on images pop up for every chapter, pulled straight from the scene descriptions and woven into an interactive reader. Text-to-speech (TTS) kicks in too, voicing the tale aloud to make it more accessible and captivating. The prototype doesn't yet handle story carryover across sessions or live text syncing with audio. It also skips built-in safety checks, pointing to age-filtering as prime territory for upgrades. All told, this setup proves a smart, streamlined way to automate rich, multimedia stories.**

**Keywords— AI-assisted storytelling, Large Language Mod-els (LLMs), Multimodal systems, Text-to-Speech (TTS), Im-age generation, Narrative generation, Content personalization, Human-computer interaction, Natural Language Processing (NLP)**

## I. INTRODUCTION

Storytelling plays a vital role in sparking kids' brainpower, emotions, and language skills. It fuels creativity, sharpens understanding, and lays the groundwork for learning. Yet most classic books and digital apps stay rigid, with little room to tweak for a child's unique tastes, skill level, or passions. This often leaves young readers bored by cookie-cutter tales that don't quite click. On top of that, parents and teachers struggle to whip up custom stories that fit a kid's world, it's a real time sink.

Thanks to breakthroughs in AI, especially large language models (LLMs) and image generators, we can now churn out smart text and matching visuals on the fly. These tools open the door to lively, made-to-order stories. Still, too many apps spit out disjointed bits, drop the plot thread, or fumble the mix of words, pictures, and sound into something unified.

This paper unveils an AI helper that builds illustrated kids' storybooks through a single, smooth pipeline. Feed it a quick prompt and chapter count, and it spits out a full package: title, chapter adventures, and illustration sketches. The magic? It crafts the whole tale in one go, keeping the storyline tight and true. Custom images then bring each chapter to life in an Identify applicable funding agency here. If none, delete this.

interactive viewer, paired with text-to-speech (TTS) for easy listening and extra fun.

The setup shines for one-off sessions but hasn't nailed ongo-ing story memory or audio-synced text glow-ups yet. Without a safety net for content filtering, kid-friendly checks top the to-do list. Ultimately, this system spotlights multimodal AI's power for fun, hands-off, personalized tales while flagging smart next steps.

### Objective

The core mission of this project is to create an AI-powered platform that produces tailored, fully illustrated children's storybooks through a unified multimodal workflow. This sys-tem streamlines every step: capturing user prompts, crafting complete

narratives, generating matching visuals, and adding voice narration all in one cohesive process. By automating these elements, it saves time for parents and educators while delivering highly engaging, coherent stories that hold kids' attention from page one.

A central aim is harnessing large language models (LLMs) to build entire multi-chapter stories from a single, straightfor-ward input like a theme or character idea. This single-pass generation ensures rock-solid narrative continuity, avoiding plot holes or jarring shifts between chapters. Outputs include a vibrant title, detailed chapter texts rich in adventure and morals, plus precise prompts for illustrations, making the whole tale feel polished and purposeful right away.

Equally important, the system uses advanced image gen-eration models to produce context-specific artwork for each chapter. These visuals aren't generic they draw directly from the story's scenes, emotions, and details (e.g., a cozy forest adventure gets lush, whimsical trees and friendly animals). They're then embedded into an interactive digital book inter-face, letting kids flip pages, zoom in, and immerse themselves visually.

For better accessibility, especially for early readers or those with visual challenges we integrate text-to-speech (TTS) to narrate the stories with natural voices, pacing, and even character-specific tones. This turns static text into a lively audio experience, perfect for bedtime or group reading.

Customization sits at the heart, letting users shape stories via inputs like child's age (for vocabulary level), interests (dinosaurs? Space?), themes (friendship, bravery), and length

(3 chapters or 10?). This personalization makes each book feel special and relevant.

Looking ahead, we spotlight critical enhancements: building in robust safety filters to scrub out any age-inappropriate language or ideas automatically, plus expanding to multi-session memory (so kids can pause and resume adventures) and interactive extras like synced text highlighting during narration or simple choice-based plot branches.

## II. LITERATURE SURVEY

To shape our AI-driven multimodal storytelling system for kids, we dug into recent research on custom tales, large lan-guage models (LLMs), and blended AI setups. These studies shed light on boosting kid engagement, streamlining story-making, and weaving together text, pictures, and sound into one captivating package.

While plenty of work tackles personalization, interactivity, or multimodal magic separately, hardly any pulls them all into a single, streamlined flow. Spotting this hole steered our design toward a more complete solution.

StoryLab: Personalized Multimodal Story Generation (Li et al., 2025) kicked things off with a teacher-led approach, using LLMs and image generators to craft stories tied to learning goals. It nails customization and education but leans on human oversight, which slows down true hands-off creation.

Metabook: AR-Based 3D Storybook Generation (Wang et al., 2025) delivers an all-in-one pipeline for dazzling 3D books via augmented reality. The immersion wows, ramping up fun, yet it demands extra gadgets and skims over deep story crafting in favor of flashy visuals.

StoryMate: LLM-Empowered Interactive Story Reading (Chen et al., 2025) dives into chat-style reading aids powered by LLMs, personalizing on the fly with real-time tweaks. Great for interaction, but it stops short of building fresh stories from scratch more helper than creator.

MM-StoryAgent: Multimodal Storybook Generation Framework (Xu et al., 2025) brings a team of AI agents to fuse text, images, and audio for rich tales. The multimodal sync shines, but the setup gets tangled with too many moving parts and synced models.

These efforts highlight a clear opening: we need a straight-forward system that nails automated full-story generation, keeps plots seamless, and blends modalities without fuss. Our approach fills that by whipping up entire multi-chapter epics in one go, then layering on scene-smart images and built-in audio delivering efficient, cohesive, and downright delightful storytelling.

## III. EXISTING SOLUTIONS

In recent years, the growing interest in personalized and interactive digital learning tools has led to the development of many AI-based storytelling systems. These platforms typically combine large language models (LLMs), image generation techniques, and user interaction to create more engaging

storytelling experiences for children. Even so, most of these systems are designed around a single strength—such as per-sonalization, interactivity, or multimodal output—rather than bringing all of them together in one fully integrated pipeline.

**StoryLab**
StoryLab is a multimodal story generation platform created mainly for educational environments, especially classrooms. It combines LLMs with image generation models to produce stories that can be adapted to learning goals and classroom needs. One of its most important features is the teacher-in-the-loop design, which helps ensure that the generated content remains appropriate and aligned with pedagogical objectives. However, because it depends on manual input and supervision, the system cannot function as a fully automated solution. This reduces its scalability and makes it less practical for independent or on-demand story generation.

**StoryMate**
StoryMate is centered on AI-supported interactive story reading, using LLMs to make the reading experience more dynamic and engaging. It allows users to interact with stories through chatbot-style conversations, personalized prompts, and guided reading support. This makes it useful for improving engagement and adapting the reading experience to individual users. However, its main purpose is to assist with reading rather than to create complete stories automatically. As a result, it does not provide multi-chapter story generation or built-in illustration generation.

**Metabook**
Metabook offers a more immersive storytelling experience by transforming stories into augmented reality ARAR-based 3D books. This approach greatly improves visualization and interaction, making the content more exciting and memo-rable for users. However, the system depends on specialized hardware such as AR devices, which limits how widely it can be used. In addition, its main focus is on presentation and visualization rather than automated story creation, which reduces its usefulness as a general-purpose storytelling system.

**MM-StoryAgent**
MM-StoryAgent proposes a multi-agent framework for gen-erating multimodal storybooks by combining text, image, and audio generation. Its staged generation process helps improve story quality and creates a richer, more immersive reading experience through synchronized modalities. However, this structure also makes the system more complex, since multiple models must be coordinated effectively. Because of this added complexity, it may not be the most efficient choice for lightweight or real-time storytelling applications.

Although these systems represent important progress in AI-driven storytelling, they still have noticeable limitations. Many rely on human intervention, require heavy computation, or de-pend on specialized hardware, while others focus only on one part of the storytelling experience. This creates a clear need

for a system that is simple, efficient, and capable of produc-ing coherent, multi-chapter, multimodal storybooks through a single unified pipeline. The proposed system addresses this need by combining structured story generation, context-aware illustration creation, and integrated audio narration in a way that supports both scalability and usability.

## IV. PROPOSED SOLUTION

Our platform offers a complete AI-powered storytelling solution that transforms a single user prompt into a fully illustrated children's storybook. Unlike other tools that zero in on just interaction, visuals, or bits of content, this one delivers a single, seamless pipeline for rock-solid narrative flow, blended text/images/audio, and full automation.

We follow a "Generate Everything First, Then Polish" strategy: a large language model (LLM) crafts the entire multi-chapter story upfront in one smooth run, guaranteeing tight logic and momentum. Next come targeted illustrations and woven-in audio. Ditching piecemeal generation cuts out plot glitches and elevates the final tale's polish.

How it stacks up against others:

- No need for human tweaks mid-process (unlike StoryLab)
- Builds whole stories from scratch, not just chat aids (unlike StoryMate)
- Runs on everyday devices, no fancy gear required (unlike Metabook)
- Keeps multimodal power simple, without agent overload (unlike MM-StoryAgent)

This makes it faster, more reachable, and ready for on-the-fly use in apps or web tools.

**How It Works**

A. Core Principles: Narrative Coherence: The full multi-chapter arc emerges in a single generation cycle, weav-ing a consistent thread no jumps, no loose ends.

Contextual Alignment: Every illustration pulls from its chapter's unique scene notes, so visuals nail the mood, action, and details perfectly.

Multimodal Integration: Text, pictures, and voice narra-tion merge into one interactive experience, like flipping digital pages with sound.

**Key Building Blocks**

User Input Handler: User Input Handler: Takes your story seed (prompt + chapter count), cleans it up, and feeds a ready prompt to kick off generation.

Story Builder: Story Builder: LLM magic: generates a complete package title, cover blurb, chapter texts, and illustration guides all structured and story-ready.

Image Creator: Image Creator: For each chapter, crafts visuals from the scene descriptions, ensuring colors, characters, and vibes sync with the words.

Book Organizer: Book Organizer: Bundles every-thing text and images into a neat digital book format, stored for instant access or sharing.

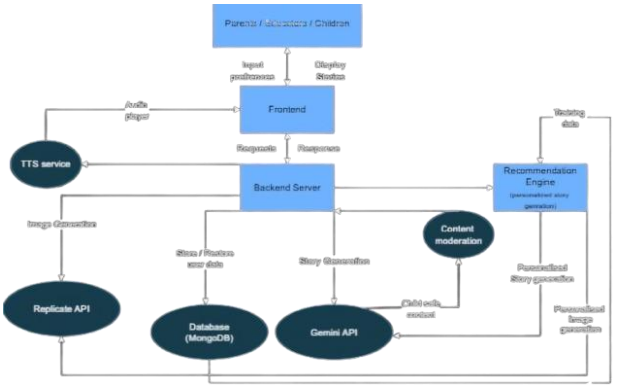


Fig. 1. System Architecture.

Interactive Reader: Interactive Reader: Displays the book with swipeable pages, plus TTS narration controls (play, pause, speed) for an engaging listen-along experience.

Think of this system like a little storytelling factory where each part has a role:

User (Parents / Educators / Children): They're the ones starting everything choosing what kind of story they want.

Frontend: This is the screen they interact with. It takes their preferences and later shows the final story, images, and audio.

Backend Server: The "brain" that handles everything behind the scenes. It receives the request, decides what to do next, and connects all the parts.

Database (MongoDB): Like memory. It stores user details, past stories, and preferences so the system can remember and improve.

Gemini API (Story Generator): The creative writer. It generates the actual story based on what the user wants.

Content Moderation: The safety checker. It makes sure the story is appropriate and child-friendly before showing it.

Recommendation Engine: The smart assistant. It learns from user behavior and suggests or creates more person-alized stories and images.

TTS Service (Text-to-Speech): The narrator. It converts the story into audio so users can listen instead of just reading.

Replicate API (Image Generation): The illustrator. It creates visuals to match the story.

**Advantages of Proposed Solutions**

End-to-End Automated Storybook Creation: From one simple prompt, the system generates a complete multi-chapter storybook containing structured text, cus-tom illustrations, and integrated audio narration. No manual assembly is required, making the process effi-cient and suitable for educators, parents, and application developers.

Strong Narrative Flow and Consistency: Since the entire story is generated in a single pass, the system ensures logical continuity, consistent character develop-ment, and smooth progression of events. This eliminates the discontinuities commonly observed in step-by-step generation approaches.

Multimodal Interactive Experience: The integration of text, illustrations, and text-to-speech (TTS) narration creates a rich and immersive storytelling environment. This enhances accessibility for early readers and im-proves engagement through a multi-sensory experience.

Accurate Text-Image Alignment: Illustrations are gen-erated based on detailed, chapter-specific descriptions, ensuring that visuals accurately represent the corre-sponding narrative. This improves comprehension and overall storytelling quality.

Scalable and Lightweight Architecture: The system is designed to operate without heavy hardware or com-plex dependencies, enabling fast performance and easy deployment across web and mobile platforms. It also supports future extensions such as personalization and safety filtering.

## V. WORKING DETAILS

At the core of the system is a structured pipeline that inte-grates story generation, image synthesis, and audio narration into a unified workflow.

Story Generation Pipeline: The process begins with user input, consisting of a story prompt and the desired number of chapters. This input is provided to a large language model (LLM), which generates the complete story in a single execution cycle. The output includes the story title, cover description, and a sequence of chapters. Each chapter contains both narrative text and a corresponding illustration description. This approach ensures strong coherence and eliminates inconsistencies found in incremental generation methods.

Illustration Generation Process: The illustration de-scriptions extracted from each chapter are used as prompts for an image generation model. These prompts are context-specific, enabling the system to produce visuals that accurately reflect the scene, characters, and emotions described in the story.

Storage and Structuring: The generated text and im-ages are organized into a structured digital book format. Each chapter is paired with its corresponding illustration and stored efficiently, allowing for easy retrieval and rendering in the user interface.

Interactive Reading and Narration: The final output is presented through an interactive reading interface. A text-to-speech (TTS) system narrates the story with user controls such as play, pause, speed adjustment, and replay. This enhances user engagement and improves accessibility for children.

## VI. CONCLUSIONS

This paper introduces an AI tool that crafts and shares illustrated kids' storybooks through a single, streamlined mul-timodal setup. Users simply drop a prompt, and it builds full multi-chapter stories with tight narrative flow, sensible pacing, and unwavering consistency. The one-shot generation sidesteps the usual pitfalls of choppy content or mismatched characters that trip up step-by-step methods.

Beyond words, it whips up spot-on illustrations for each chapter, drawn right from the scene details to bridge text and visuals perfectly. This amps up understanding and fun. Top it off with text-to-speech (TTS) for easy audio playback, and you've got a setup that welcomes early readers and beyond.
Setting it apart from others, our design delivers full au-tomation without the hassle of tangled agents, extra gadgets, or human hand-holding. It's built to scale, run fast, and reach everyone.

The prototype sticks to one-and-done sessions, but it spot-lights ripe upgrades: carrying stories over multiple sittings, glowing text synced to audio, and smart filters for kid-safe vibes. In the end, this shows multimodal AI's power to reinvent storytelling making it personal, immersive, and ready to roll, while boosting creativity, learning, and real connections.

## REFERENCES

1. X. Xu et al., "MM-StoryAgent: Immersive Narrated Storybook Video Generation with a Multi-Agent Paradigm across Text, Image and Audio," arXiv preprint arXiv:2503.05242, 2025. A247, pp. 529–551, April 1955.
2. Z. Li et al., "StoryLab: Empowering Personalized Learning for Children Through Teacher-Guided Multimodal Story Generation," Springer, 2025.
3. J. Chen et al., "Characterizing LLM-Empowered Personalized Story-Reading and Interaction for Children," CHI Conference, 2025.
4. Y. Wang et al., "Metabook: A Mobile-to-Headset Pipeline for 3D Story Book Creation in Augmented Reality," arXiv:2405.13701, 2025.
5. J. Kim et al., "A Multi-Modal Story Generation Framework with AI-Driven Storyline Guidance," Electronics, 2023..
6. Z. Lin et al., "Narratology Meets Text-to-Image: Consistency in AI-Generated Storybook Illustrations," Artificial Intelligence Review, 2026.
7. E. Bensaid et al., "FairyTailor: A Multimodal Generative Framework for Storytelling," arXiv:2108.04324, 2021.