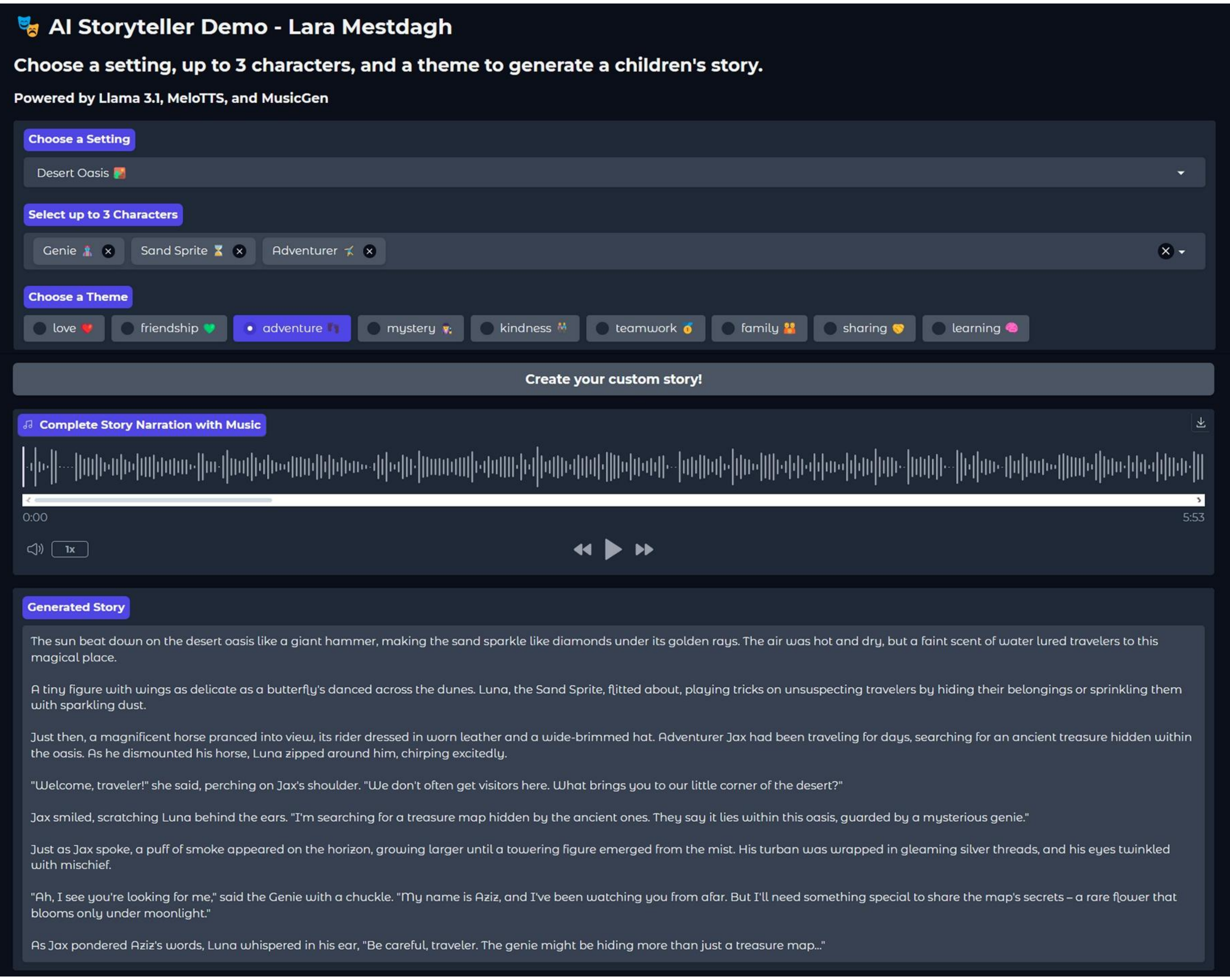

AI Storyteller Demo - Lara Mestdagh

Choose a setting, up to 3 characters, and a theme to generate a children's story.

Powered by Llama 3.1, MeloTTS, and MusicGen

Choose a Setting: Desert Oasis

Select up to 3 Characters: Genie, Sand Sprite, Adventurer

Choose a Theme: ○ love ○ friendship ● adventure ○ mystery ○ kindness ○ teamwork ○ family ○ sharing ○ learning

Create your custom story!
Complete Story Narration with Music
0:00
5:53
1x

Generated Story

The sun beat down on the desert oasis like a giant hammer, making the sand sparkle like diamonds under its golden rays. The air was hot and dry, but a faint scent of water lured travelers to this magical place.

A tiny figure with wings as delicate as a butterfly's danced across the dunes. Luna, the Sand Sprite, flitted about, playing tricks on unsuspecting travelers by hiding their belongings or sprinkling them with sparkling dust.

Just then, a magnificent horse pranced into view, its rider dressed in worn leather and a wide-brimmed hat. Adventurer Jax had been traveling for days, searching for an ancient treasure hidden within the oasis. As he dismounted his horse, Luna zipped around him, chirping excitedly.

"Welcome, traveler!" she said, perching on Jax's shoulder. "We don't often get visitors here. What brings you to our little corner of the desert?"

Jax smiled, scratching Luna behind the ears. "I'm searching for a treasure map hidden by the ancient ones. They say it lies within this oasis, guarded by a mysterious genie."

Just as Jax spoke, a puff of smoke appeared on the horizon, growing larger until a towering figure emerged from the mist. His turban was wrapped in gleaming silver threads, and his eyes twinkled with mischief.

"Ah, I see you're looking for me," said the Genie with a chuckle. "My name is Aziz, and I've been watching you from afar. But I'll need something special to share the map's secrets – a rare flower that blooms only under moonlight."

As Jax pondered Aziz's words, Luna whispered in his ear, "Be careful, traveler. The genie might be hiding more than just a treasure map..."

HOW CAN AI BE USED TO GENERATE AND NARRATE A CHILDREN'S STORY FOR AGES 5–8 WITH ACCOMPANYING BACKGROUND MUSIC?

INTERNAL PROMOTOR: WOUTER GEVAERT
EXTERNAL PROMOTOR: WOUTER GROVE

RESEARCH QUESTION CARRIED OUT BY

LARA MESTDAGH

FOR OBTAINING THE BACHELOR'S DEGREE IN THE

MULTIMEDIA & CREATIVE TECHNOLOGIES

HOWEST | 2024-2025

Preface

This thesis marks the completion of my bachelor's degree in Multimedia and Creative Technologies in the AI Engineer minor at Howest University of Applied Sciences in Kortrijk, Belgium. The work presented here explores the central research question:

"How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?"

Although this thesis is not directly related to my internship in South Africa, I received valuable support from both the internship company and external experts. I chose this topic because it allowed me to explore AI technologies that were previously unfamiliar to me, such as story generation and music synthesis.

The development of this project benefited from input from individuals from the University of the Western Cape (UWC), including professionals in software development, artificial intelligence, and education. I also interviewed a coworker from the internship company to reflect on how the research could translate into practical, real-world use cases.

The thesis begins by discussing the research and the technical demo developed during the "Research Project" module from semester 5 of the MCT curriculum. The project involved building an AI-powered pipeline that generates short children's stories, converts them into narrated audio and adds musical transitions to create an engaging storytelling experience.

After explaining the proof of concept, I critically reflect on the results and their potential value in educational and media applications, including recommendations for developers and researchers interested in similar AI-driven solutions.

I sincerely thank everyone who supported me throughout this thesis.

At Howest, I am grateful to my internal mentor, Wouter Gevaert, for his consistent guidance, feedback, and encouragement during this final project.

From the internship company Data4, I would like to thank Rudolf Visser, a data engineer and analyst, for taking the time to participate in an interview and for providing practical insights into the applicability of my research.

I also extend my gratitude to the professionals from the University of the Western Cape who generously shared their expertise during the reflection phase:

- Dr. Gassant Gamiet, faculty of education and president of the Coding and Robotics Club
- André Daniels, digital media coordinator at the Centre for Innovative Education & Communication Technologies (CIECT)
- Dr. Wouter Grove, manager of the Future Innovation Lab

Their thoughtful input helped me evaluate the broader relevance of this project, and I wholeheartedly appreciate their openness and time.

Lara Mestdagh – 30/05/2025

Abstract

This bachelor’s thesis investigates how artificial intelligence can generate child-friendly multimedia stories that combine text, speech narration, and transition music clips for early childhood education. Addressing the shortage of engaging and accessible digital content for young learners, the project develops a modular AI pipeline that produces complete story experiences from minimal user input. The system integrates open-source models for story generation, text-to-speech synthesis, and music composition coordinated through a Gradio-based interface.

To ensure age-appropriate output for children aged 5–8, the pipeline uses structured prompt engineering, automated readability scoring, and audio quality validation. A metadata-driven approach lets users define story parameters, such as setting, characters, and themes, guiding the generation process across all pipeline parts.

Experts in education, software development, and AI evaluated the prototype through interviews. They emphasized its potential for early learning environments and praised the integration of multiple AI methods. At the same time, they pointed out limitations, including output variability and sensitivity to prompts. Based on expert feedback, this thesis recommends improving the system’s robustness, language and cultural localization, and accessibility.

Overall, the project demonstrates the promise of AI that integrates text, narration, and music for creative educational tools and outlines a pathway for scalable, child-safe story generation.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
AD	Action Discriminator
AI	Artificial Intelligence
API	Application Programming Interface
AR/VR	Augmented Reality / Virtual Reality
BPM	Beats Per Minute
CALT	Centre for African Language Teaching
CAPS	Curriculum Assessment Policy Statements
CIECT	Centre for Innovative Education and Communication Technologies
CPU	Central Processing Unit
ETL	Extract, Transform, Load
FairytaleQA	Fairytale Question Answering Dataset
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPU	Graphics Processing Unit
GROVE	Retrieval-augmented Complex Story Generation Framework
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
LLM	Large Language Model
MCD	Mel-Cepstral Distortion
MCT	Multimedia and Creative Technologies
MIDI	Musical Instrument Digital Interface
MOS	Mean Opinion Score
MR	Mixed Reality
POPIA	Protection of Personal Information Act
REST API	Representational State Transfer Application Programming Interface

RLHF	Reinforcement Learning from Human Feedback
RMS	Root Mean Square
SPSS	Statistical Parametric Speech Synthesis
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
SUPERB	Speech Processing Universal PERformance Benchmark
SWAG	Storytelling With Action Guidance
TTS	Text-To-Speech
TTSDS	Text-To-Speech Distribution Score
UI/UX	User Interface / User Experience
UWC	University of the Western Cape
WAV	Waveform Audio File Format
WER	Word Error Rate
XR	Extended Reality

Table 1: List of abbreviations.

Glossary

Term	Definition
Amplitude	The height of a sound wave, related to loudness.
Artificial intelligence	Technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.
Back-end	The server-side part of the application handling logic, processing, and integration of AI models.
Cloud deployment	Hosting the application on cloud infrastructure for scalability and accessibility.
Contextual chaining	Passing outputs from one generation step as context for subsequent steps, improving coherence in multi-part generation tasks.
Data4	The internship company during my international internship in South Africa.
Diffusion model	A generative AI model that refines noise into structured output, used for music and image generation.
Digital literacy	The ability to confidently and critically use digital technologies to find, evaluate, create, and communicate information.
Docker	A platform for developing, shipping, and running applications inside containers.
Extended reality	An umbrella term that encompasses VR, AR, and MR, referring to immersive technologies that blend digital and physical environments for interactive experiences.
Few-shot learning	A training method where a model learns to perform a task using only a few examples.
Flesch reading ease score	A readability metric that rates text on a 100-point scale, higher scores indicate easier reading.
Flesch-Kincaid grade level	A readability test that assigns a numerical value to a text, indicating the U.S. school grade level at which a reader can understand it.
Front-end	The user-facing part of the application where users interact with the system.
Gradio	An open-source Python library for building browser-based user interfaces for machine learning models.
Hierarchical transformer	A neural network architecture that processes data at multiple levels of abstraction, often used in advanced LLMs and music models.
In-context learning	A method where a model learns to perform a task by being given examples directly in the input prompt, without updating its weights.
Librosa	A Python package for music and audio analysis.
Licensing	Legal permissions and restrictions governing the use and distribution of software or models.

Localization	Adapting content or systems to suit the language, culture, and preferences of a specific target audience or region.
Mel-spectrograms	A visual representation of sound frequencies over time, scaled to match human pitch perception.
Metadata	Input elements such as setting, characters, and theme used to guide story generation in this project.
Modular architecture	System design where components (modules) can be independently developed, replaced, or upgraded without affecting the whole system.
NumPy	A Python library for numerical computing, used for processing audio arrays.
Ollama	A platform for running large language models locally.
Onboarding	The process of guiding new users through setup and initial interaction with a system, often designed to improve understanding and usability.
Open-source	Software whose source code is made freely available for modification and distribution.
Personalization	The customization of content or experience based on user preferences or characteristics, such as age, language, or interests.
Phoneme	The smallest unit of sound in speech, used in TTS systems for accurate pronunciation.
Phygital	Describes experiences or products that blend physical and digital elements, often used in the context of XR and immersive media.
Prompt engineering	The process of designing and refining input prompts to guide the output of AI models, especially LLMs.
Proof of concept	A prototype or demonstration project built to validate technical feasibility before developing a full product.
Prosody	The rhythm, stress, and intonation of speech, important for expressive and natural-sounding TTS.
Sample rate	The number of samples of audio carried per second, measured in Hz or kHz.
Scipy	A Python library used for scientific and technical computing, including audio processing.
Self-supervised learning	A machine learning method where the model learns from unlabeled data by predicting parts of the input.
Spatial computing	The use of AI, computer vision, and XR to blend virtual and physical environments, enabling 3D interactions and context-aware digital experiences.
Spectral accuracy	A measure of how closely a generated or processed audio signal matches the original in terms of its frequency content.
Storyboarding	The process of visually outlining the sequence of a narrative or user experience, commonly used in film, animation, and interactive media to plan story structure and user flow.

Symbolic music generation	A method of music generation based on MIDI or note-level representations instead of audio waveforms.
Textstat	A Python library to calculate statistics from text. It provides utilities for calculating text metrics with built-in functions to help determine readability, complexity, and grade level.
Three-act structure	A narrative framework that divides stories into setup, conflict, and resolution.
Tokenization	The process of breaking text into smaller units (tokens) for processing by language models.
Upsampling	Increasing the sample rate of an audio file to match another, ensuring compatibility during audio processing.
Vocoder	A component in a TTS system that converts spectrograms into waveform audio, enabling the synthesis of human-like speech.

Table 2: Glossary.

1 Introduction

1.1 Inspiration

This project draws inspiration from early childhood experiences of learning to read, particularly the comfort and excitement of bedtime stories and reading alongside an adult. These formative moments are reimagined in a modern context through the development of an AI-powered storytelling tool. The goal is to recreate the immersive and engaging nature of narrated stories by combining automated story generation, expressive narration, and thematic music.

To personalize the storytelling experience, users can guide the creation process by selecting a story setting, main characters, and a central theme. These choices influence the direction and content of the generated story, ensuring both structure and creativity while keeping the narrative appropriate and engaging for young listeners.

The topic was selected for its alignment with interests in artificial intelligence and narrative media. The project builds on a lifelong appreciation of stories that spark imagination and emotion, aiming to combine this narrative richness with modern AI techniques. Technologies such as story generation, text-to-speech synthesis, and music composition contribute to a compelling and accessible storytelling experience for children.

1.2 Goals

This thesis is guided by the central research question:

"How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?"

This interdisciplinary research integrates natural language generation, speech synthesis, and AI-based music composition. The aim is to explore the potential of automated storytelling as both a technical system and a creative tool, with particular attention to its value in educational and entertainment contexts.

To support this exploration, a series of sub-questions have been defined that examine the system's individual components, their integration, and the overall impact on user experience and reading engagement.

The seven sub-questions addressed in this thesis are as follows:

1. *“What AI techniques are suitable for generating child-friendly stories appropriate for children aged 5–8?”*
2. *“What are the technical challenges in combining AI-based story generation, narration, and music composition into a single, efficient application, and how can they be addressed?”*
3. *“How can we ensure that the generated stories are free from inappropriate content and remain engaging for early readers?”*
4. *“How can short AI-generated music transitions enhance the storytelling experience in children's stories, and what techniques improve their thematic consistency?”*
5. *“How can TTS technology be optimized and evaluated to ensure that the narration is engaging, clear, and age-appropriate?”*
6. *“What results emerge when using the application to support children’s reading skills, and what conclusions can be drawn from these results?”*
7. *“How can feedback, both during development and through a user-facing feedback function, be used to improve the AI models and the overall storytelling experience?”*

1.3 Tools and methodology

This research follows a two-phase approach, beginning with a theoretical study and then developing a technical proof of concept. The theoretical phase consists of a literature review focused on recent advancements in AI-driven story generation, text-to-speech synthesis, and music composition. Building on these insights, the second phase focuses on creating a modular application that combines these technologies into a unified and interactive system.

The application is fully developed in Python, using Visual Studio Code as the primary development environment. Gradio creates a browser-based interface that allows users to interact with the system easily by selecting story parameters such as the setting, characters, and theme.

The storytelling pipeline integrates the following pre-trained open-source models:

- Llama 3.1 (Meta AI) via Ollama generates stories from structured prompts using a three-act narrative format.
- MeloTTS produces natural-sounding speech from the generated story text via a Dockerized REST API.
- MusicGen-Small (Meta AI) via the HuggingFace transformers library generates short music clips to transition between story sections.

These models are combined into a step-by-step pipeline. The process starts with story generation, proceeds through narration, and concludes with the integration of background music. Instead of relying on custom-trained models, the system leverages prompt engineering and validation logic to ensure output quality and thematic consistency.

The codebase is structured into components responsible for story generation, narration, music composition, audio merging, and interface control. Generated outputs, including story text, narration audio, and background music, are stored in separate directories, which are excluded from version control. A dedicated licenses folder is included to ensure complete transparency regarding the use of third-party models.

1.4 Structure

This thesis is divided into eight chapters, structured to follow the full research and development process from concept to implementation and evaluation:

- Chapter 1 – Introduction: Introduces the background and context of the project, including the central research question, supporting sub-questions, the chosen methodology, the tools used, and the overall structure of the thesis.
- Chapter 2 – Research: Reviews relevant literature and technologies related to AI-driven storytelling, TTS narration, and music generation. This chapter addresses research that extends beyond the standard MCT curriculum.
- Chapter 3 – Technical Research: Describes how the storytelling application was developed, with attention to the modular architecture, the individual generation steps, validation processes, and the final interface.
- Chapter 4 – Reflection: Provides a critical evaluation of the project outcomes, supported by insights from interviews with academic and industry professionals.
- Chapter 5 – Recommendations: Offers practical suggestions for developers, educators, and researchers working on similar AI-based storytelling tools. These recommendations are grounded in both the technical results and external feedback.
- Chapter 6 – Conclusion: Summarizes the key findings from previous chapters and directly answers the central research question.
- Chapter 7 – References: Lists all sources cited throughout the thesis, formatted in IEEE style.
- Chapter 8 – Appendices: Contains additional materials that support the thesis, including transcripts of expert interviews, reports from guest lectures, as well as installation and user manuals for the application.

2 Research

2.1 General

To support the development of an AI-powered storytelling application, a theoretical study was conducted to explore the core technologies involved. This research phase focused on identifying suitable approaches for story generation, TTS narration, and music composition, three distinct yet interrelated fields within artificial intelligence.

While each of these domains has seen significant progress in recent years, fully integrated systems that combine all three into a cohesive storytelling pipeline remain uncommon. This chapter examines how existing AI models and methods can be adapted and combined to serve the project's central goal: producing engaging, safe, and age-appropriate stories for children aged 5–8.

Each research area corresponds to a specific stage of the storytelling workflow. The story generation section focuses on techniques that produce logically coherent and imaginative narratives. TTS research explored how modern systems deliver expressive, intelligible narration. Music generation analysis investigates techniques for creating thematically consistent background audio that enhances the storytelling experience.

The findings presented in this chapter are based primarily on academic research papers, open-source model documentation, and technical blog articles from leading AI research groups. Together, these studies lay the foundation for the design and implementation decisions made during the technical development phase, which is discussed in Chapter 3.

2.2 Story generation

Recent developments in AI-driven storytelling have led to the emergence of models capable of generating coherent and contextually appropriate narratives. While early systems were rule-based and heavily scripted, modern approaches rely on large language models (LLMs) trained on vast corpora of text. These models are now capable of generating longer-form stories with a level of fluency and thematic consistency that was previously unattainable.

However, significant challenges remain. As highlighted by Isozaki [1], AI-generated stories often suffer from issues such as logical inconsistencies, repetitive phrasing, abrupt pacing, or weak narrative arcs. To address these limitations, more structured methods were introduced, including metadata-guided prompts, iterative refinement, and evaluation mechanisms that prioritize coherence and engagement.

Two notable frameworks, GROVE [2] and SWAG [3], demonstrate different approaches to improving story quality. GROVE enhanced generation by retrieving and integrating relevant evidence, as shown in Figure 1. SWAG, on the other hand, used an action-guidance mechanism, known as the action discriminator (AD), to steer story development in a logical and engaging direction (see Figure 2). Interestingly, studies showed that the SWAG approach, when paired with smaller open-source models such as Llama-2 or Mistral-7B, could outperform much larger models like GPT-3.5 in human evaluation benchmarks.

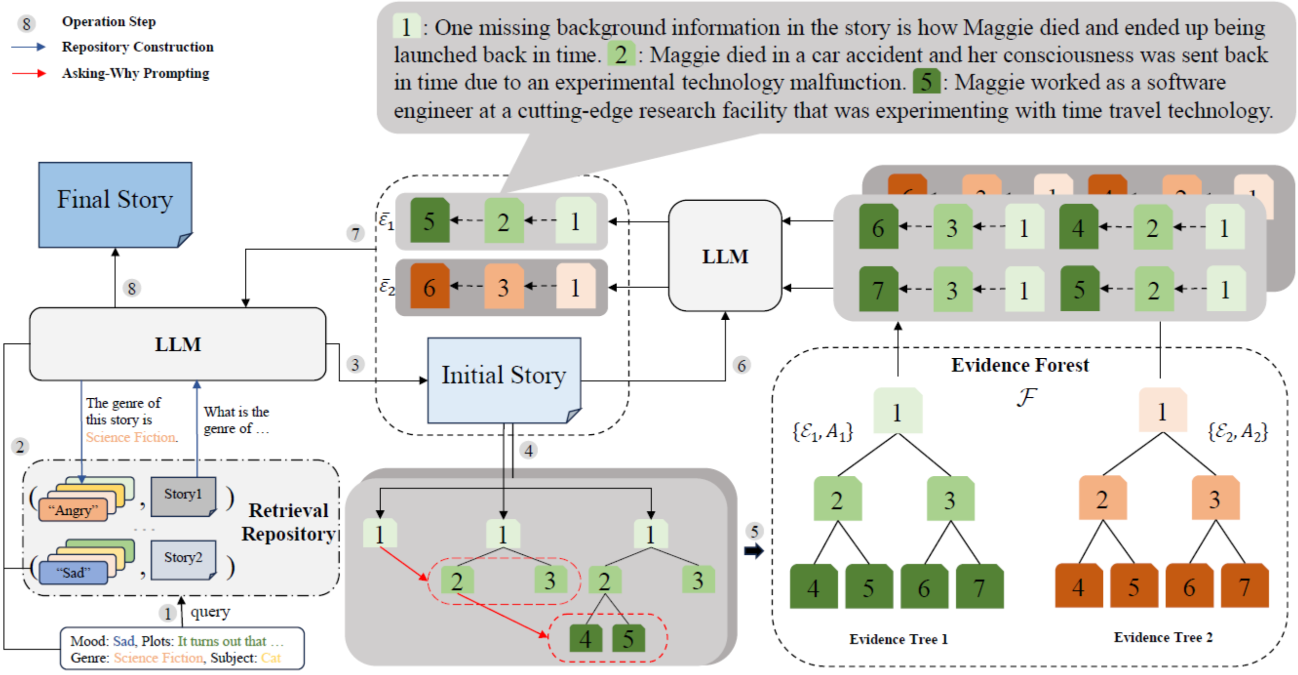

8 Operation Step
Repository Construction
Asking-Why Prompting
1 : One missing background information in the story is how Maggie died and ended up being launched back in time. 2 : Maggie died in a car accident and her consciousness was sent back in time due to an experimental technology malfunction. 5 : Maggie worked as a software engineer at a cutting-edge research facility that was experimenting with time travel technology.
Final Story
LLM
Initial Story
LLM
Evidence Forest
$\mathcal{F}$
$\{\mathcal{E}_1, A_1\}$
$\{\mathcal{E}_2, A_2\}$
Evidence Tree 1
Evidence Tree 2
$\mathcal{E}_1$
$\mathcal{E}_2$
The genre of this story is Science Fiction.
What is the genre of ...
"Angry"
Story1
"Sad"
Story2
Retrieval Repository
query
Mood: Sad, Plots: It turns out that ...
Genre: Science Fiction, Subject: Cat

Figure 1: GROVE architecture for story generation, adapted from [2].

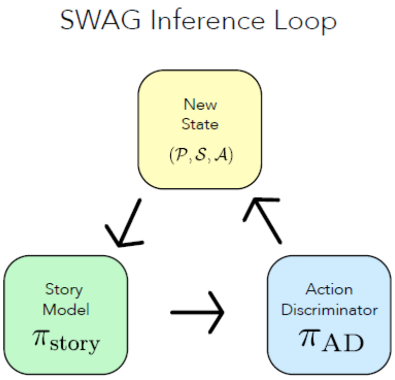


Figure 2: The SWAG inference loop, adapted from [3].

Despite these improvements, several studies have identified persistent shortcomings in AI storytelling, particularly in terms of consistency, character development, and thematic depth. One such study, *Evaluating Creative Short Story Generation in Humans and Large Language Models* [4], found that while language models can produce grammatically correct stories, the outputs often lack originality and narrative depth when compared to human-authored texts. The generated content tends to be predictable and overly reliant on common patterns seen in training data. As shown in Figure 3, readability scores for LLM-generated texts were often comparable to or even higher than those of human-written content, yet this does not necessarily correlate with narrative quality or creativity.

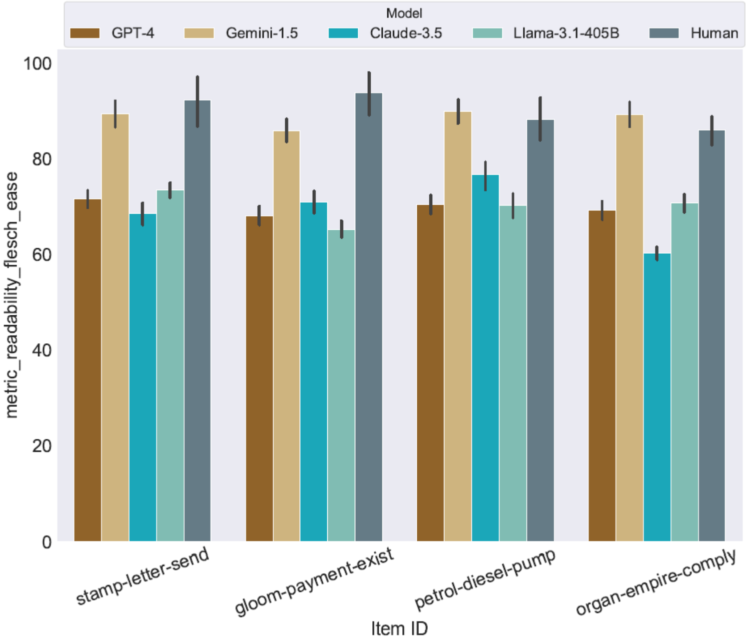
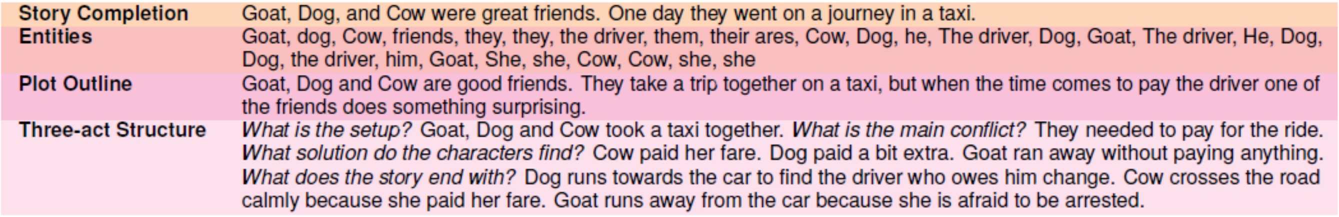


Figure 3: Results for lexical complexity metric measured by readability using Flesch reading ease score, copied from [4].

Similarly, *Assessing Language Models' Worldview for Fiction Generation* [5] emphasizes that current models frequently struggle to maintain a consistent fictional universe. This often leads to contradictions in plot or character behavior, undermining overall coherence and believability. These limitations point to a broader challenge in creative generation: LLMs tend to replicate surface-level structure without a deeper understanding of narrative logic or progression.

One approach to improving narrative coherence is the use of structured storytelling techniques, such as the three-act structure explored in *Crosslingual Story Planning and Generation with Large Language Models* [6]. This framework divides a story into three segments: setup, conflict, and resolution, providing a natural arc that guides both character development and plot progression. When integrated into AI systems, the structure helps maintain logical flow and clarity throughout the narrative. An example of how this structure is applied in planning AI-generated stories is shown in Figure 4, where key questions guide the model through each phase of the narrative.



Story Completion	Goat, Dog, and Cow were great friends. One day they went on a journey in a taxi.
Entities	Goat, dog, Cow, friends, they, they, the driver, them, their ares, Cow, Dog, he, The driver, Dog, Goat, The driver, He, Dog, Dog, the driver, him, Goat, She, she, Cow, Cow, she, she
Plot Outline	Goat, Dog and Cow are good friends. They take a trip together on a taxi, but when the time comes to pay the driver one of the friends does something surprising.
Three-act Structure	*What is the setup?* Goat, Dog and Cow took a taxi together. *What is the main conflict?* They needed to pay for the ride. *What solution do the characters find?* Cow paid her fare. Dog paid a bit extra. Goat ran away without paying anything. *What does the story end with?* Dog runs towards the car to find the driver who owes him change. Cow crosses the road calmly because she paid her fare. Goat runs away from the car because she is afraid to be arrested.

Figure 4: Story plan examples. Questions in italics correspond to main events in three-act structure, copied from [6].

Evaluating the quality of these stories remains a significant challenge. The FairytaleQA dataset [7] offers a structured framework for assessment, focusing on aspects such as character actions, emotional arcs, and causal relationships. Additionally, *Do Language Models Enjoy Their Own Stories?* [8] explores automated evaluation strategies, examining how LLMs can rate stories based on coherence, relevance, and engagement. Although these approaches show promise for scalable evaluation, human judgment remains essential, particularly for assessing deeper narrative consistency, creativity, and thematic richness.

Sub-question context

Collectively, these studies form the foundation for the design choices made in the technical implementation described later in this thesis. In particular, the findings from this research address the first sub-question:

“What AI techniques are suitable for generating child-friendly stories appropriate for children aged 5–8?”

The literature indicates that combining large language models with structured storytelling frameworks, such as the three-act structure, and guided prompting techniques is effective in producing content suitable for young readers. Approaches like retrieval-augmented generation (GROVE) and action-based planning (SWAG) help improve narrative flow and coherence, while datasets like FairytaleQA enable structured evaluation of story content. Together with careful prompt design and validation logic, these techniques form a promising foundation for generating engaging, age-appropriate stories.

2.3 Narration generation

Text-to-speech (TTS) technology has seen substantial advancements in recent years, enabling the automatic conversion of written text into natural-sounding speech. TTS plays a vital role in applications such as virtual assistants, audiobooks, accessibility tools, and AI-generated media content [9]. For this project, TTS was a critical component of the storytelling pipeline, providing narrated audio output for the generated stories and contributing significantly to the overall user experience.

Traditionally, TTS systems consist of three main components: text analysis, acoustic modeling, and vocoding. As illustrated in Figure 5, these stages worked together to process written text into spoken output. Early systems used rule-based methods for text analysis, handling tokenization, phoneme conversion, and syntactic structure, but have since been largely replaced by neural models such as Char2Wav and DeepVoice, which learn direct mappings from characters to speech [9].

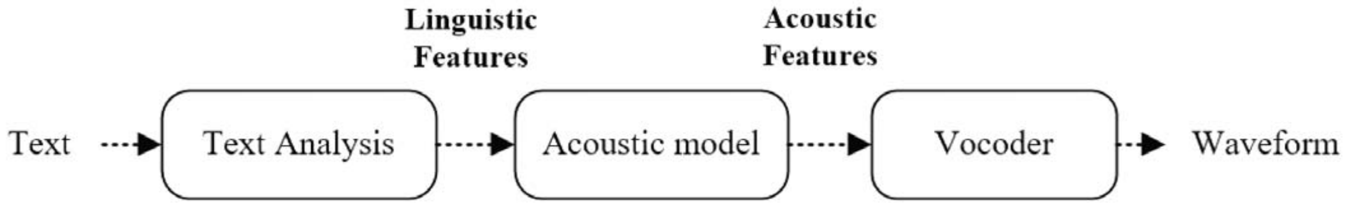


Figure 5: General structure of TTS systems, copied from [9].

In the acoustic modeling phase, linguistic features are transformed into intermediate representations like Mel-spectrograms. Deep learning models such as Tacotron, FastSpeech, and their successors have significantly improved speech naturalness and prosody, outperforming earlier methods like concatenative synthesis and statistical parametric synthesis (SPSS) [10]. As shown in Figure 6, modern architectures like FastSpeech 2 combined prosody prediction and waveform generation into a unified structure, improving efficiency and synthesis quality.

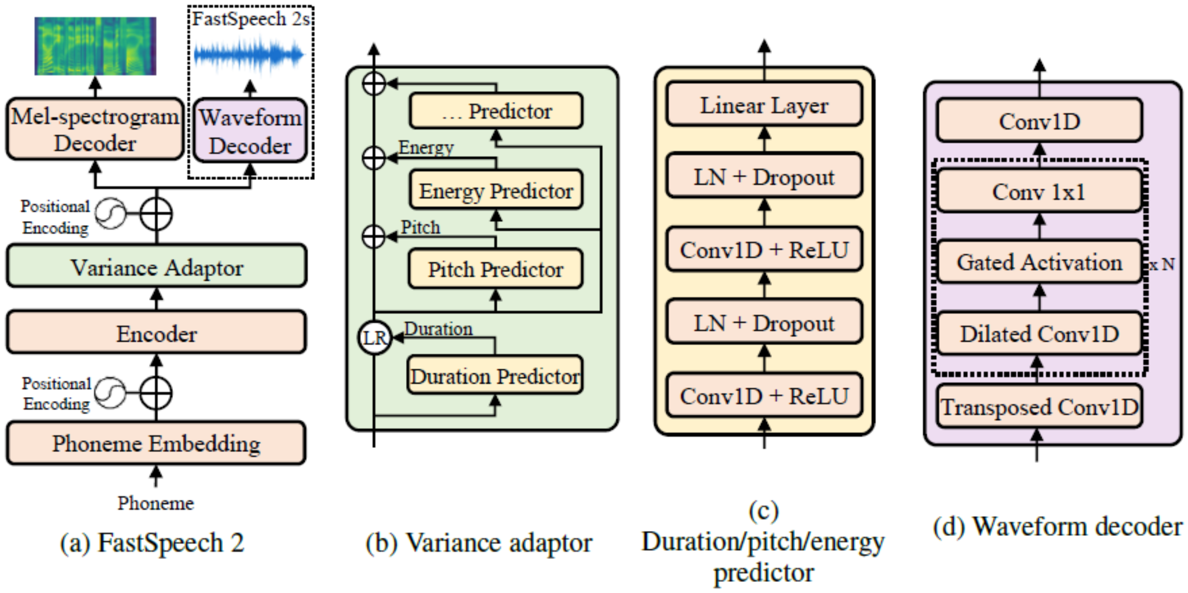


Figure 6: The overall architecture for FastSpeech 2, adapted from [10].

For waveform generation, neural vocoders such as WaveNet, WaveGlow, and LPCNet have replaced older approaches like STRAIGHT and WORLD, offering lifelike, expressive speech suitable for end-to-end pipelines [9]. Recent models such as Tacotron 2, Deep Voice 3, and FastSpeech 2 have further streamlined the generation process by integrating all components into unified architectures [10].

Beyond performance, adaptability has also improved. Newer systems such as NaturalSpeech 3 incorporate in-context and few-shot learning, enabling them to work with unfamiliar voices with minimal training data. This flexibility was particularly valuable for applications where voice diversity is important. Figure 7 illustrates how the system scales data and synthesizes voices efficiently [11].

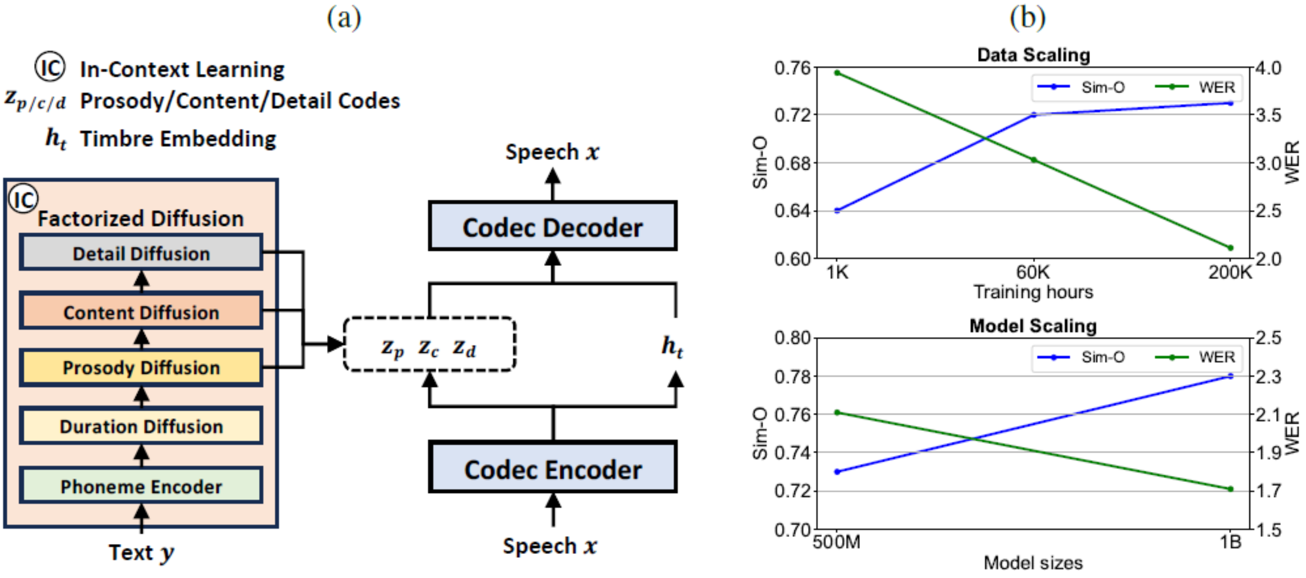


Figure 7: (a) Overview of the NaturalSpeech 3 system. (b) Model and data scaling efficiency, adapted from [11].

Evaluating TTS output remains an essential part of model development. Common methods include Mean Opinion Score (MOS) for subjective evaluation, Word Error Rate (WER) for intelligibility, and Mel-Cepstral Distortion (MCD) for spectral accuracy. More comprehensive benchmarks, like SUPERB, began to assess models across multiple dimensions, including speaker similarity, fluency, and prosody [12].

Despite these advances, challenges remain in achieving expressive, emotionally resonant narration, qualities particularly important in storytelling. To address this, models like SpeechT5 proposed a joint pretraining architecture that unifies speech and text embeddings via cross-modal vector quantization. This approach allowed for greater control over prosody, tone, and expressive variation. The model architecture is shown in Figure 8 [13].

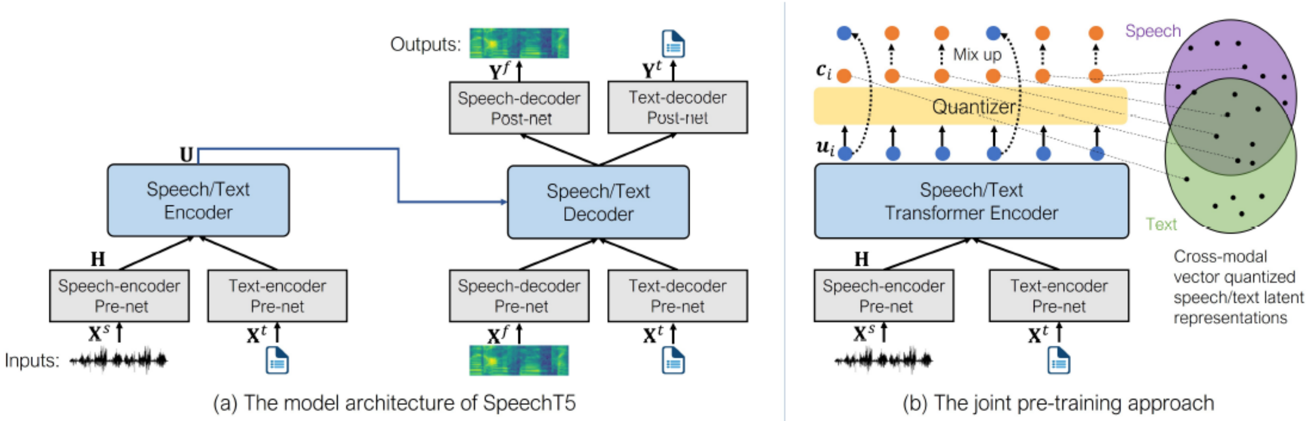


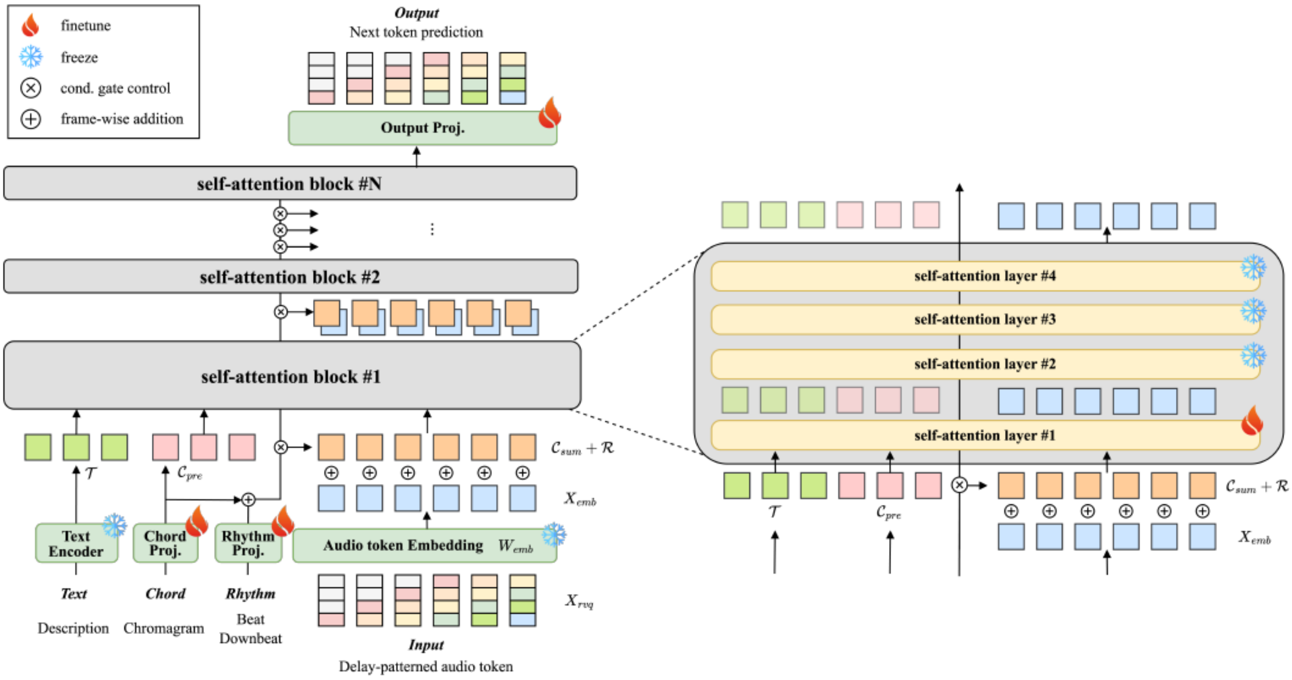
Figure 8: The model architecture of SpeechT5, adapted from [13].

Overall, modern TTS systems are well-suited for integration into AI storytelling tools. When paired with effective quality control methods, these models can produce narration that is clear, engaging and well-suited for young audiences.

2.4 Music generation

Text-to-music generation has evolved rapidly in recent years, driven by deep learning architectures capable of translating textual descriptions into musical compositions. Advancements in transformers, diffusion models, and symbolic representations improved the quality, coherence, and control of AI-generated music. Despite this progress, challenges in maintaining long-term structure, enabling expressive control, and achieving computational efficiency remain.

One widely used method treats music as a sequence of tokens that are analogous to natural language. MusicLM [14], for example, applied hierarchical transformers and contrastive text-music embeddings to generate music that adhered to a desired style and maintained consistency over time. MusiConGen [15] expanded on this approach by introducing explicit control over rhythm and harmony, enabling structured compositions rather than freeform generation. The underlying architecture of MusiConGen, shown in Figure 9, highlights how self-attention mechanisms and explicit rhythm and chord control are used to guide structure in the generated music.



(a) MusiConGen model structure

(b) self-attention block

Figure 9: The model structure of MusiConGen and the self-attention block, adapted from [15].

An alternative approach involved the use of diffusion models, which iteratively refined noise into structured audio. Models such as ERNIE-Music [16] and Moûsai [17] demonstrated strong capabilities in producing high-quality waveforms, especially for ambient or cinematic audio. However, they often lacked fine-grained structural control and demanded substantial computational resources.

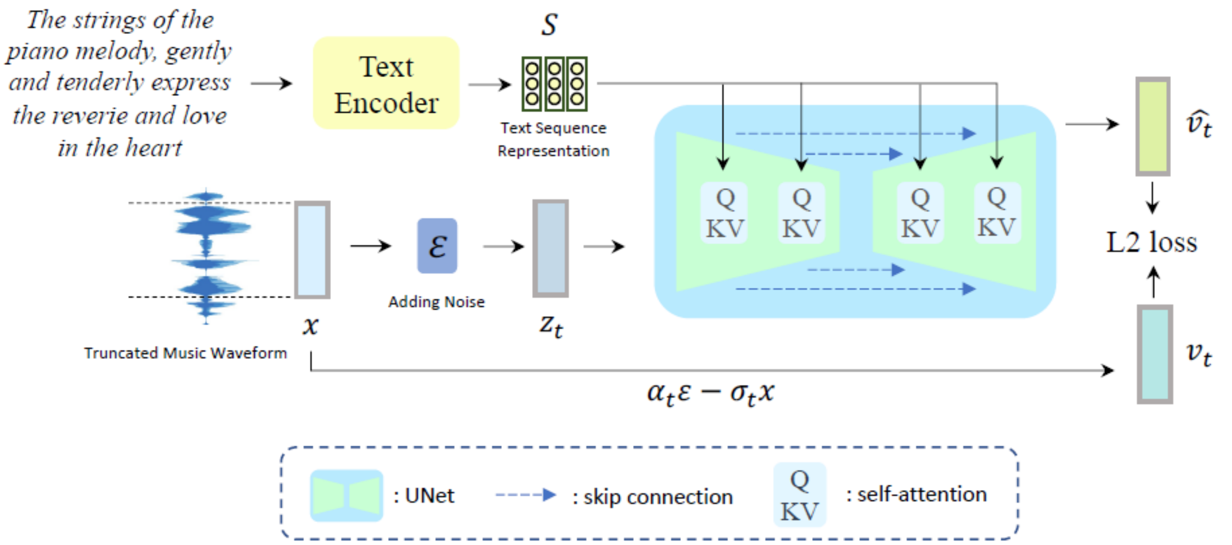

The strings of the piano melody, gently and tenderly express the reverie and love in the heart
Text Encoder
S
Text Sequence Representation
Q KV
$\hat{v}_t$
L2 loss
v_t
Truncated Music Waveform
x
$\mathcal{E}$
Adding Noise
z_t
$\alpha_t \varepsilon - \sigma_t x$
: UNet
: skip connection
: self-attention

Figure 10: The overall architecture of text-to-music generation training, adapted from [16].

Some systems adopted a symbolic representation strategy, relying on MIDI-based formats rather than raw audio. MusicFlow [18], for instance, applied flow-matching techniques to enhance long-term coherence and temporal control. Symbolic approaches offer greater structural precision but often at the cost of the expressive richness found in audio-based generation. Figure 10 shows a full training pipeline that combines text encoding, sequence representation, and waveform generation. This structure reflects how current models attempt to bridge semantic and acoustic representations through joint learning.

Despite their strengths, several common limitations persist across current models:

- Long-term structure: Most systems generate only short clips (typically 3–30 seconds) and struggle to preserve musical coherence over longer compositions.
- Controllability: Pure text prompts lack specificity in defining musical elements such as tempo, chord progression, or instrumentation. MusiConGen addressed this by introducing BPM and chord progression control.
- Computational efficiency: Diffusion-based models offered impressive audio realism but remained computationally expensive, posing challenges for real-time or low-resource use cases.
- Data limitations: Many systems relied on self-supervised learning due to the scarcity of high-quality, labeled text-music datasets.

Evaluation of AI-generated music also revealed subjective limitations. Compared to human-composed works, the music was often perceived as less emotionally expressive and less structured over extended timeframes. Listener preferences varied by genre and use case, with generated music showing promise in ambient or background scenarios while facing more difficulty in producing melodic or rhythmically complex content.

Ethical and legal concerns also emerged. Since many models are trained on copyrighted music, the legal status of generated compositions remains unclear. Questions about licensing, authorship, and the commercial use of AI music were still being debated at the time of writing.

Future research directions included improving dataset quality, refining hybrid models that combine symbolic and audio features, and developing real-time music generation systems. These advancements could increase the practicality of AI music in storytelling applications, particularly in interactive media, educational content, and audio-enhanced books.

3 Technical Research

3.1 System architecture

The AI storytelling application was built as a local proof of concept, structured around a modular pipeline that separates user interaction from back-end processing. The application consists of two main layers:

- Front-end: A Gradio-based interface used for metadata selection and output delivery.
- Back-end pipeline: A sequence of Python modules responsible for story generation, narration, music generation, and audio combination.

The modular architecture is shown in the system flowchart (Figure 11), which illustrates each component from user input to audio output.

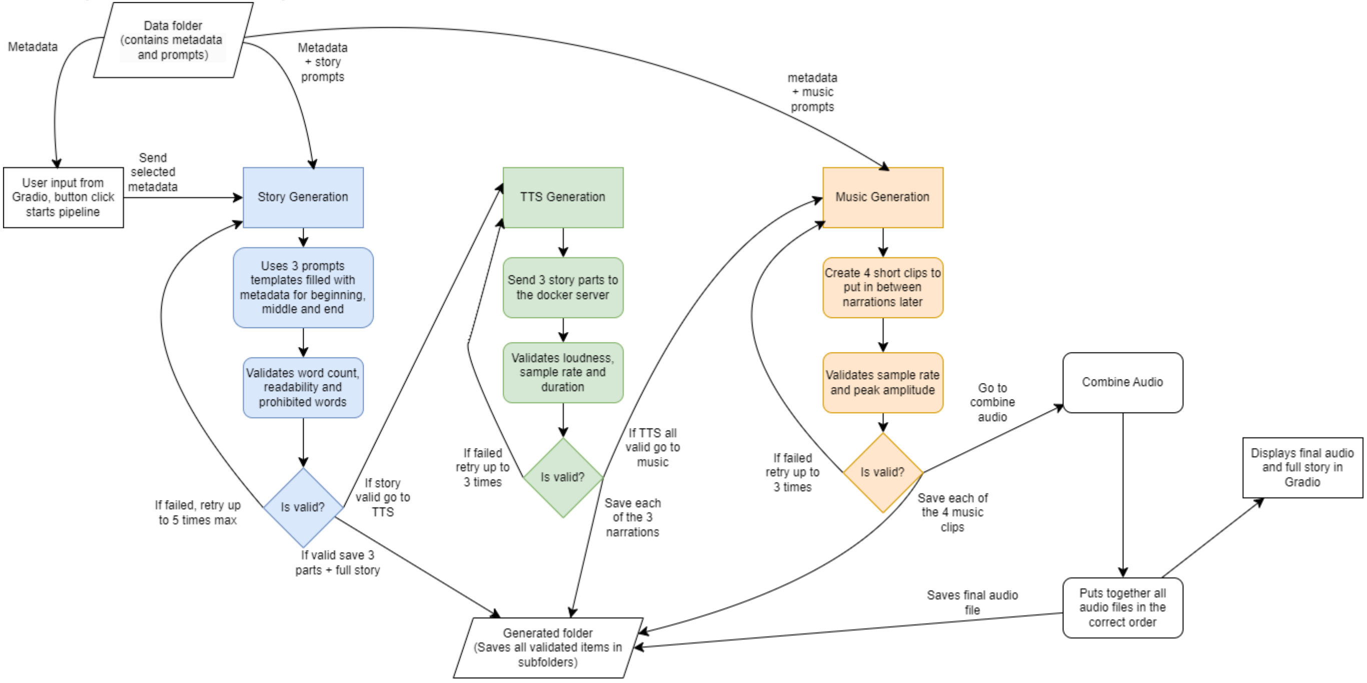

Data folder (contains metadata and prompts)
Metadata
Metadata + story prompts
metadata + music prompts
User input from Gradio, button click starts pipeline
Send selected metadata
Story Generation
Uses 3 prompts templates filled with metadata for beginning, middle and end
Validates word count, readability and prohibited words
Is valid?
If failed, retry up to 5 times max
If story valid go to TTS
If valid save 3 parts + full story
TTS Generation
Send 3 story parts to the docker server
Validates loudness, sample rate and duration
Is valid?
If failed retry up to 3 times
If TTS all valid go to music
Save each of the 3 narrations
Music Generation
Create 4 short clips to put in between narrations later
Validates sample rate and peak amplitude
Is valid?
If failed retry up to 3 times
Go to combine audio
Save each of the 4 music clips
Combine Audio
Puts together all audio files in the correct order
Saves final audio file
Displays final audio and full story in Gradio
Generated folder (Saves all validated items in subfolders)

Figure 11: Application flowchart.

The pipeline was composed of the following steps:

1. Metadata input (via Gradio front-end): Users interacted with a custom Gradio UI to select a story setting, one to three compatible characters, and a central theme. Character availability was dynamically filtered based on the selected setting to maintain internal narrative logic.
2. Story generation (LLM module): The back-end used the selected metadata to generate a story in three stages, beginning, middle, and end, using prompt templates based on a three-act structure. The story was created using the Llama 3.1 model running locally via Ollama and was automatically validated and regenerated if needed before moving to the next step.
3. Narration (TTS module): Each story segment was converted into audio using MeloTTS. Audio outputs were validated and regenerated as needed to ensure proper format and quality.
4. Music generation (text-to-music module): Four short instrumental transitions were generated using Facebook’s MusicGen-Small model, influenced by the selected setting and a predefined instrument list. Like narration, music clips were also validated before use and were regenerated if needed.
5. Audio assembly: The validated narration and music clips were merged into a single audio file using the Python libraries of Librosa and Scipy. The structure included brief silences for pacing and followed a consistent order:
   Intro music → Beginning narration → Transition 1 → Middle narration → Transition 2 → End narration → Outro music.
6. Output display (via Gradio front-end): The final audio and full story text were returned to the user for playback and reading in the browser.

Integrating these stages into a single application presented several challenges. Running local models required balancing resource usage and speed. Despite model optimizations, a full run typically took 8–10 minutes for each story. An overview of estimated generation times is provided in Table 3, based on testing with an Intel i5-9300H CPU and a Nvidia GeForce RTX 2060 GPU.

Pipeline step	**Time**
Story generation	7 – 8 minutes
Narration generation	1 – 2 minutes
Music generation	1 – 2 minutes

Table 3: Time per generation for each pipeline step.

Sub-question context

Each module was developed as a self-contained unit with its own validation and retry logic. This modular approach allowed for automatic error handling and simplified testing and upgrades. For instance, a new music or TTS model could be swapped in without affecting the rest of the application. The modular design also leaves room for future extensions, such as adding new story settings, expanding music clip durations, or supporting multiple narrator voices.

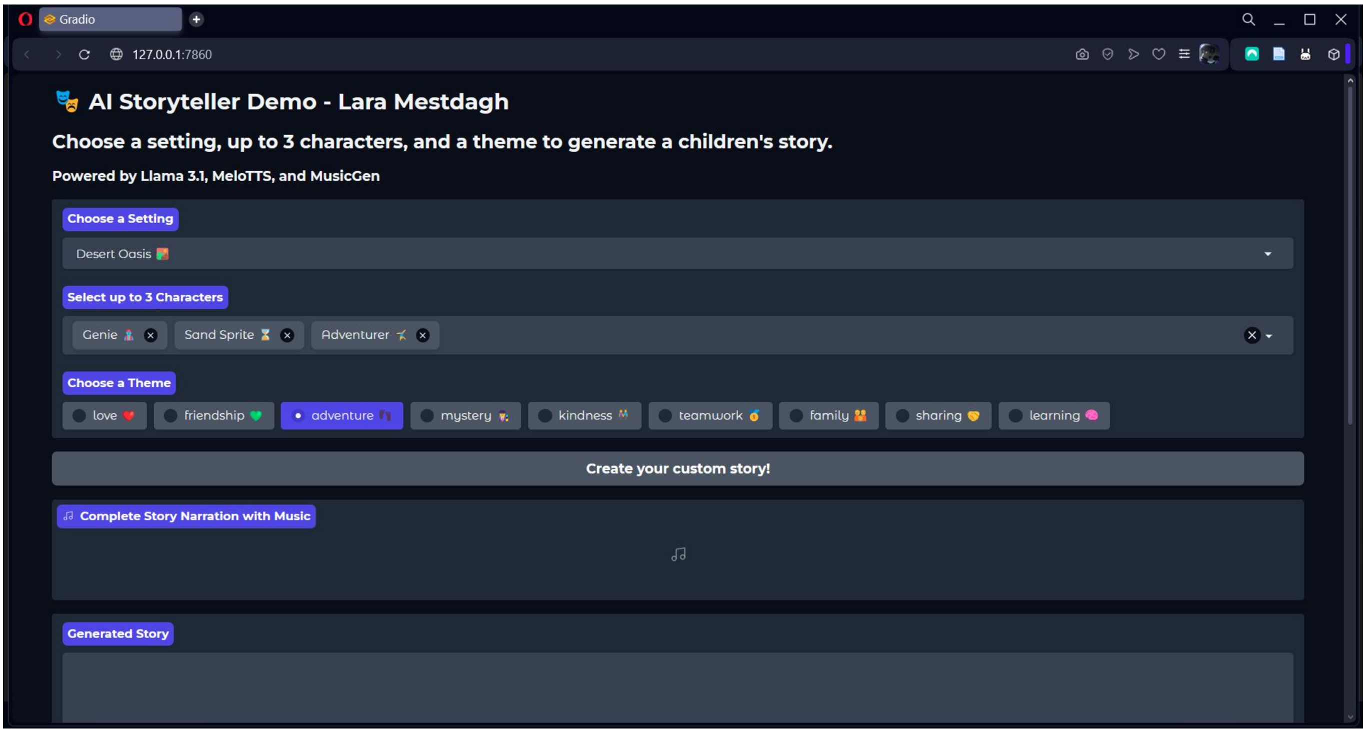
This structure directly addressed the technical sub-question:

"*What are the technical challenges in combining AI-based story generation, narration, and music composition into a single, efficient application, and how can they be addressed?"*

By separating responsibilities, streamlining model integration, and ensuring independent validation per stage, the system achieved both reliability and flexibility within a complex, AI-driven workflow.

3.2 Meta-driven story creation

The story generation process began with user-selected metadata, namely the setting, characters, and theme, collected via the Gradio interface. The system dynamically filtered character options based on the chosen setting to ensure logical consistency (e.g., mermaids were only available in the underwater world). Themes, in contrast, remained independent and could be applied to any setting-character combination, offering creative flexibility. An example of the metadata selection interface is shown in Figure 12.



Gradio

127.0.0.1:7860

AI Storyteller Demo - Lara Mestdagh

Choose a setting, up to 3 characters, and a theme to generate a children's story.

Powered by Llama 3.1, MeloTTS, and MusicGen

Choose a Setting

Desert Oasis

Select up to 3 Characters

Genie ⊗ Sand Sprite ⊗ Adventurer ⊗

Choose a Theme

○ love ○ friendship ● adventure ○ mystery ○ kindness ○ teamwork ○ family ○ sharing ○ learning

Create your custom story!

Complete Story Narration with Music

Generated Story

Figure 12: Gradio UI, metadata selection before generation.

The back-end generated stories using Meta's Llama 3.1 8B model, run locally via Ollama. This model was selected after a testing phase in which several alternatives were explored, including Phi-4 and Qwen 2.5. While Phi-4 showed strong coherence and Qwen 2.5 offered creativity, Llama 3.1 struck the best balance between story quality, inference speed, and resource efficiency, making it the most suitable choice for this project's goals and constraints.

Prompt engineering and iterative development

In the first implementation, the entire story was generated in one pass using a single prompt. However, this approach often produced short, structurally flat results with limited character progression. To overcome these issues, the pipeline was restructured using a three-act storytelling method, which split the generation into:

- Beginning: Introduce the setting, characters, and core problem.
- Middle: Expand the conflict or challenge and build narrative tension.
- Ending: Resolve the story and deliver a satisfying conclusion.

Each section was generated separately using a dedicated, fine-tuned prompt. The prompts included specific instructions to:

- Keep the language simple and age-appropriate.
- Encourage short, clear sentences.
- Add playful moments and interactions.
- Reinforce the selected theme organically.

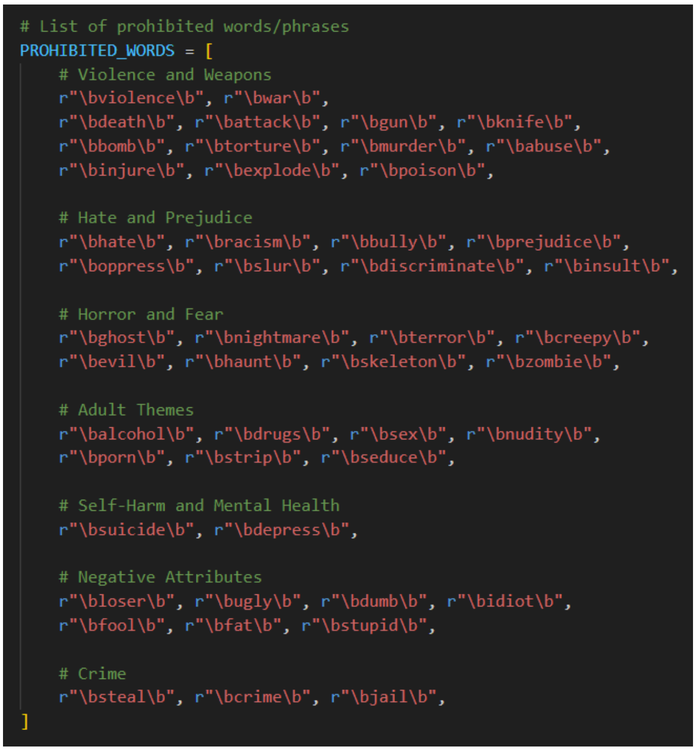
The system passed previously generated sections as context for the next, enabling character development and story flow to evolve logically. These prompts were iteratively refined during development based on output quality, length, and thematic clarity.

Validation pipeline and retry logic

After generating the full story (beginning, middle, and end), it was automatically validated against a set of strict quality and safety criteria:

- Readability: Targeted a Flesch Reading Ease score of 75–100.
- Grade Level: Ensured content complexity remained equal to or below the 6th-grade level.

- Length: Enforced a word count between 800–1300 words.
- Content Safety: Checked against a predefined prohibited words list designed to filter out inappropriate language, references to violence, adult content, and negative stereotypes. A screenshot of this list can be seen in Figure 13.



```
# List of prohibited words/phrases
PROHIBITED_WORDS = [
    # Violence and Weapons
    r"\bviolence\b", r"\bwar\b",
    r"\bdeath\b", r"\battack\b", r"\bgun\b", r"\bknife\b",
    r"\bbomb\b", r"\btorture\b", r"\bmurder\b", r"\babuse\b",
    r"\binjure\b", r"\bexplode\b", r"\bpoison\b",

    # Hate and Prejudice
    r"\bhate\b", r"\bracism\b", r"\bbully\b", r"\bprejudice\b",
    r"\boppress\b", r"\bslur\b", r"\bdiscriminate\b", r"\binsult\b",

    # Horror and Fear
    r"\bghost\b", r"\bnightmare\b", r"\bterror\b", r"\bcreepy\b",
    r"\bevil\b", r"\bhaunt\b", r"\bskeleton\b", r"\bzombie\b",

    # Adult Themes
    r"\balcohol\b", r"\bdrugs\b", r"\bsex\b", r"\bnudity\b",
    r"\bporn\b", r"\bstrip\b", r"\bseduce\b",

    # Self-Harm and Mental Health
    r"\bsuicide\b", r"\bdepress\b",

    # Negative Attributes
    r"\bloser\b", r"\bugly\b", r"\bdumb\b", r"\bidiot\b",
    r"\bfool\b", r"\bfat\b", r"\bstupid\b",

    # Crime
    r"\bsteal\b", r"\bcrime\b", r"\bjail\b",
]
```

Figure 13: Prohibited words list.

If any check failed, the system discarded the story and retried generation, up to five attempts per request. This validation logic was embedded in the back-end code, ensuring that only stories that met all standards were saved and forwarded to the narration step.

Sub-question context

This multi-step approach directly addressed the research sub-question:

"How can we ensure that the generated stories are free from inappropriate content and remain engaging for early readers?"

By combining strict prompt engineering, modular generation, and comprehensive automatic validation, the system consistently produced stories that were imaginative, structurally sound, and safe for young audiences. The use of sequential prompting with contextual chaining also improved logical flow and character depth, resulting in stories that were both entertaining and appropriate for an early childhood audience.

3.3 Narration with TTS

Once the story passed all validation checks, it was forwarded to the text-to-speech (TTS) module for audio narration. This phase was handled by MeloTTS, a lightweight, open-source TTS model chosen for its fast inference time, high-quality voice output, and ease of local deployment via a Dockerized REST API.

To maintain narrative pacing and structural alignment, each of the three story sections, beginning, middle, and end, was narrated separately. Segmenting the narration to match the story's act-based format allowed for better control over playback flow and audio assembly in later stages.

TTS integration, tuning, and validation

Each section of text was sent to the MeloTTS API, which returned an audio file in WAV format. Before finalizing any output, each narration clip was validated using Librosa, which checked for:

- Loudness (RMS) ensures clarity and presence.
- Duration to avoid cutoffs or incomplete segments.
- Sample Rate consistency at 44.1 kHz.

Any clip that failed validation was automatically regenerated up to three times until it met all required standards.

In addition to this back-end integration, tuning decisions were made to optimize the voice output for young listeners. Several voice and accent options were tested using the online HuggingFace demo for MeloTTS. The clearest and most child-appropriate voice was selected based on articulation and tone, in this case, the British-accented female voice. Furthermore, the narration speed was adjusted to 0.9x, slightly slowing the delivery to enhance understanding for the 5–8 age group without sounding unnatural. This combination of structural segmentation, audio validation, and fine-tuned output made the narration both reliable and accessible.

Sub-question context

This setup directly addressed the technical research sub-question:

"*How can TTS technology be optimized and evaluated to ensure that the narration is engaging, clear, and age-appropriate?*"

By integrating an efficient and customizable TTS engine, applying rigorous validation, and carefully selecting the voice profile and speech tempo, the system ensured that narration was both engaging and developmentally appropriate. The result was a smooth and consistent audio experience aligned with the linguistic and auditory needs of the target age group.

3.4 Music generation and transitions

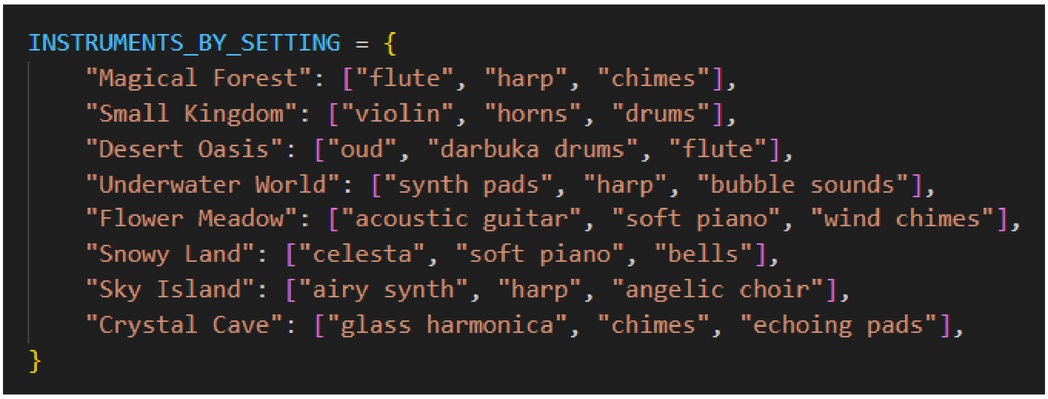
The original plan was to generate continuous background music to play softly throughout the narrated story. However, due to current limitations in AI-generated music models, particularly around length, structure, and compute constraints, this approach proved impractical. Most models, including those evaluated for this project, generate relatively short audio clips (typically under 30 seconds) and often lack the long-term coherence needed to support full narrative scenes.

As a solution, the design is adapted to use short musical transitions instead. These clips act as ambient markers between story sections, supporting the story's pacing and atmosphere without requiring full-scene coverage. In total, four music clips, each approximately three seconds long, were generated for every story:

- Intro music (before narration begins)
- Transition 1 (between beginning and middle)
- Transition 2 (between middle and end)
- Outro music (after the story concludes)

Model selection and prompt design

The transitions were generated using MusicGen-Small by Meta, a text-to-music model capable of producing short instrumental clips from descriptive prompts. Each prompt was customized based on the selected setting and paired with a predefined list of instruments that reflected the environment's tone and aesthetic. A screenshot of this list can be seen in Figure 14.



```
INSTRUMENTS_BY_SETTING = {
    "Magical Forest": ["flute", "harp", "chimes"],
    "Small Kingdom": ["violin", "horns", "drums"],
    "Desert Oasis": ["oud", "darbuka drums", "flute"],
    "Underwater World": ["synth pads", "harp", "bubble sounds"],
    "Flower Meadow": ["acoustic guitar", "soft piano", "wind chimes"],
    "Snowy Land": ["celesta", "soft piano", "bells"],
    "Sky Island": ["airy synth", "harp", "angelic choir"],
    "Crystal Cave": ["glass harmonica", "chimes", "echoing pads"],
}
```

Figure 14: Instrument list by setting.

These instrument mappings were defined in the application's configuration and automatically injected into the music prompt templates. This ensured that each story's soundscape was thematically consistent and easily adaptable to new settings in the future.

Audio compatibility and validation

Music clips were generated at a default sample rate of 32 kHz, as supported by MusicGen-Small. However, since the narration audio was processed at 44.1 kHz, all music clips were upsampled using Librosa to ensure consistency during the final audio assembly.

Each music file was also validated automatically before use. The validation included checks on the following:

- Amplitude: To prevent inaudible or silent clips.
- Sample Rate: Matched the 44.1 kHz from TTS generation to ensure audio consistency.

Clips that failed validation were discarded and regenerated up to three times per request, ensuring quality without disrupting the pipeline.

Sub-question context

This approach addressed the research sub-question:

"*How can short AI-generated music transitions enhance the storytelling experience in children's stories, and what techniques improve their thematic consistency?*"

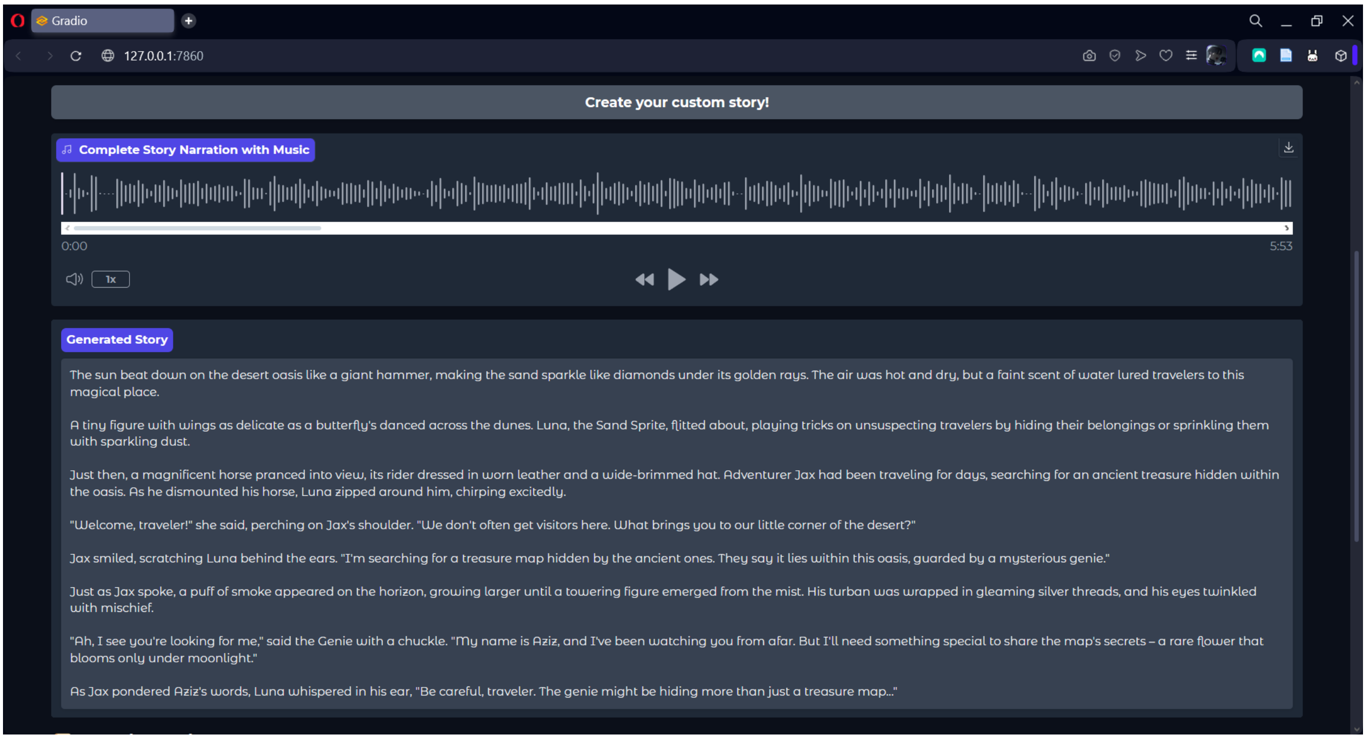
By moving from continuous background music to shorter, strategically placed transitions, the system maintained immersion without sacrificing quality or reliability. The use of setting-specific instrument lists and prompt-driven generation ensured that the music enhanced, rather than distracted from, the storytelling experience. These short cues effectively reinforced emotional tone and narrative structure, making the overall experience more engaging for young listeners.

3.5 Audio synchronization and assembly

The final step of the pipeline combined the validated narration and music clips into a single synchronized audio file. This was handled by the audio assembly module, which ensured that all clips aligned in timing, format, and sample rate.

Each story followed a fixed playback sequence: *Intro music → Beginning narration → Transition 1 → Middle narration → Transition 2 → End narration → Outro music*

Short silences were inserted to improve pacing and allow natural breathing space between segments. This also helped smooth transitions and avoided abrupt changes in tone or volume. Once aligned, all segments were concatenated using NumPy arrays and processed into a normalized WAV file via Scipy. Peak amplitude was checked and adjusted if necessary to prevent audio clipping and ensure clean playback across devices. The result was a fully synchronized audio story that was playable directly within the Gradio interface. An example of this can be found in Figure 15.


Gradio
127.0.0.1:7860
Create your custom story!
Complete Story Narration with Music
0:00
5:53
1x
Generated Story

The sun beat down on the desert oasis like a giant hammer, making the sand sparkle like diamonds under its golden rays. The air was hot and dry, but a faint scent of water lured travelers to this magical place.

A tiny figure with wings as delicate as a butterfly's danced across the dunes. Luna, the Sand Sprite, flitted about, playing tricks on unsuspecting travelers by hiding their belongings or sprinkling them with sparkling dust.

Just then, a magnificent horse pranced into view, its rider dressed in worn leather and a wide-brimmed hat. Adventurer Jax had been traveling for days, searching for an ancient treasure hidden within the oasis. As he dismounted his horse, Luna zipped around him, chirping excitedly.

"Welcome, traveler!" she said, perching on Jax's shoulder. "We don't often get visitors here. What brings you to our little corner of the desert?"

Jax smiled, scratching Luna behind the ears. "I'm searching for a treasure map hidden by the ancient ones. They say it lies within this oasis, guarded by a mysterious genie."

Just as Jax spoke, a puff of smoke appeared on the horizon, growing larger until a towering figure emerged from the mist. His turban was wrapped in gleaming silver threads, and his eyes twinkled with mischief.

"Ah, I see you're looking for me," said the Genie with a chuckle. "My name is Aziz, and I've been watching you from afar. But I'll need something special to share the map's secrets – a rare flower that blooms only under moonlight."

As Jax pondered Aziz's words, Luna whispered in his ear, "Be careful, traveler. The genie might be hiding more than just a treasure map..."

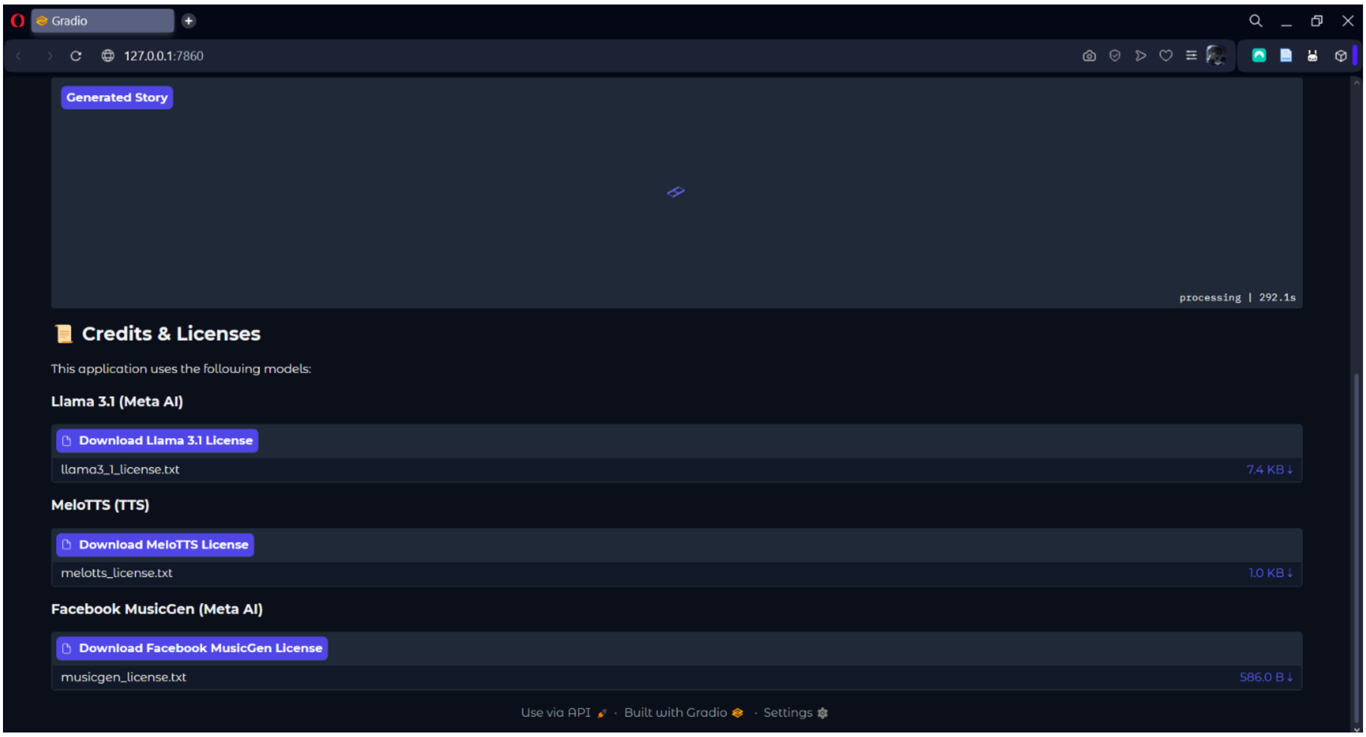
Figure 15: Gradio UI after full generation.

3.6 Interactive front-end and output

The final component of the system was the Gradio interface, which provided a guided and intuitive way for users to interact with the storytelling pipeline. Users selected their story parameters, triggered story generation, and received an audio-visual output if all validation steps passed.

The front-end displayed both the full story text and a built-in audio player for playback. If any part of the generation process failed validation (e.g., inappropriate content or unreadable output), an error message was shown, and the user was prompted to retry.

The interface design was iteratively improved during development. Visual elements like emojis were added to enhance usability, and validation messages were implemented to provide users with immediate feedback in case of missing selections or generation errors. A license section was also added to the bottom of the interface, where users can download the model licenses as separate TXT files. This can be seen in Figure 16.


Gradio
127.0.0.1:7860
Generated Story
processing | 292.1s

📜 Credits & Licenses

This application uses the following models:

Llama 3.1 (Meta AI)

Download Llama 3.1 License

llama3_1_license.txt 7.4 KB ↓

MeloTTS (TTS)

melotts_license.txt 1.0 KB ↓

Facebook MusicGen (Meta AI)

Download Facebook MusicGen License

musicgen_license.txt 586.0 B ↓

Use via API · Built with Gradio · Settings

Figure 16: Gradio UI, credit and licenses.

4 Reflection

4.1 Approach and expert involvement

This chapter offers a critical reflection on the development and potential of the AI storytelling proof of concept. Rather than focusing solely on technical execution, it evaluates the project's practical value, strengths, limitations, and future potential in real-world educational and technological contexts. The goal is not only to assess the technical viability of the tool but also to understand its broader relevance for stakeholders in education, technology, and digital inclusion.

Multiple experts were consulted to ensure the reflection was well-grounded rather than purely self-referential. Their feedback helps validate core assumptions, question initial design choices, and identify key opportunities for improvement. This external validation process is crucial in assessing how well the tool aligns with real-world needs and in contextualizing the findings beyond the initial prototype phase. This process also deepens understanding of how the tool could evolve into a scalable, adaptable solution for real-world use.

The interviewees include:

- Dr. Gassant Gamiet, faculty of education specializing in science education and technology, also the president of the Coding and Robotics Club at the University of the Western Cape (UWC)
- André Daniels, digital media coordinator at the Centre for Innovative Education & Communication Technologies (CIECT), UWC
- Dr. Wouter Grove, manager of the Future Innovation Lab, UWC
- Rudolf Visser, data engineer and analyst at Data4

Their insights are woven throughout the chapter, shaping both the structure and content of this reflection. Each expert brings a unique perspective, representing different facets of the project's potential impact, from classroom application to systemic deployment at scale. This reflection contributes directly to answering the main research question by evaluating how effectively the proposed AI storytelling tool meets real-world educational and societal needs. It also provides a foundation for future research and development, drawing from both practical findings and professional experience.

4.2 Validation process and external perspectives

To ensure a robust and valuable evaluation of the AI storytelling proof of concept, feedback has been gathered from professionals with expertise in education, digital media, and software development. The goal is to critically assess its relevance, usability, and limitations from diverse perspectives. This multidisciplinary input is essential not only for identifying potential flaws but also for revealing improvements that may not have been apparent during initial development.

Interviewees were chosen to represent both technical and educational domains. On the technical side, Rudolf Visser and Dr. Wouter Grove provide insights into the system's architecture, modularity, and deployment potential. From an educational perspective, Dr. Gassant Gamiet and André Daniels offer feedback on the project's literacy support, creative value, and cultural suitability. Their combined expertise allows for a nuanced critique that considers both implementation feasibility and educational effectiveness or impact on learning.

Each participant received a full information packet ahead of the interview, including:

- A narrated demo video showcasing the application in use
- Multiple stories and narration outputs
- The full user manual and installation guide
- A link to the GitHub repository with the full source code and documentation

The interviews were conducted one-on-one and focused on both technical and educational questions. Topics discussed included: storytelling structure, narration and music quality, user interface accessibility, implementation feasibility, and ethical risks. All interviews were recorded and transcribed for accurate analysis, and the transcripts are included in Chapter 8. Interview questions were tailored to the participant's domain of expertise, allowing for specific feedback on elements ranging from system architecture to child engagement and safety protocols.

The external validation process tested both the system's technical robustness and the assumptions made during development against the experiences of professionals in education and technology. For example, Dr. Grove emphasizes the importance of safety guardrails and content validation for young users, while André Daniels highlights the need for personalization and warns against over-reliance on AI in creative contexts. Dr. Gamiet speaks of the tool's potential in supporting early literacy and the importance of cultural adaptability, and Rudolf Visser offers feedback on modularity, scalability, and deployment challenges. Their observations often confirmed internal findings while also introducing fresh perspectives, such as preparing offline-friendly versions and establishing user feedback loops, areas that had not been a primary focus during development.

Their feedback forms the foundation of the critical reflection that follows, helping to distinguish between theoretical ambitions and practical, real-world considerations. By incorporating these external insights, the project gains not only credibility but also a clearer roadmap for future development.

4.3 Key insights: strengths, weaknesses, and real-world alignment

Technical and educational strengths

A consistent theme across all interviews is the project's modular architecture. Rudolf Visser and Dr. Wouter Grove both highlight the clear separation between story generation, narration, and music as a major technical strength. This modularity is seen as a future-proof design, allowing for easy swapping or upgrading of individual modules (such as improved TTS or music models) without requiring a complete system overhaul. Visser notes that this design facilitates adaptation for different deployment scenarios, such as cloud-based platforms or mobile apps, and can support the integration of new features as technology evolves. This modular structure also allows contributors to improve specific elements without altering the entire system. This flexibility is especially valuable for future visual or interactive features, making it well-suited to open-source development.

From an educational standpoint, Dr. Gassant Gamiet and André Daniels emphasize the tool's potential to support early literacy. They praise the structured storytelling approach, age-appropriate vocabulary, and the option for children to hear narration aloud. These features help develop listening comprehension, vocabulary, and narrative awareness in young children. The ability to customize stories by setting, character, and theme is also regarded as a way to foster creative thinking and engagement. The customizable metadata structure, a core innovation, is praised for its pedagogical value, creating stories that can match a child's context, interests, and learning goals. This flexibility is especially important in multicultural classrooms, where standard content may not always resonate.

Identified weaknesses and limitations

Despite these strengths, all interviewees identify significant limitations. The current user interface is described as too basic for independent use by children or teachers without technical support. Daniels and Dr. Gamiet specifically note the absence of visual aids, onboarding guidance, or touch-friendly controls, elements essential for early learners and non-technical users. In its current form, the interface functions well as a testing environment but requires significant redesign for widespread use in real-world learning scenarios. The lack of intuitive navigation, particularly for pre-literate users, limits the tool's accessibility and user engagement.

Music integration is another area flagged for improvement. While the use of three-second transitions is a creative workaround given current model constraints, interviewees feel this is insufficient for conveying emotion or enhancing immersion. They recommend exploring dynamic background scoring in future versions as model capabilities improve. Additionally, the current TTS voices are seen as serviceable but not expressive enough for storytelling, an element that could impact the emotional tone and listener connection.

A further concern is the limited diversity in generated content. Without careful prompt engineering, stories often default to Western-centric settings and character archetypes. Both Dr. Gamiet and Daniels warn that this can hinder cultural relevance and inclusivity, especially in multilingual or diverse classroom contexts. They advocate for localized prompts, support for underrepresented languages, and explicit cultural adaptation to ensure broader applicability. While the current system technically supports localization, the actual output content does not yet reflect the diversity required for global or regional deployment. This disconnect between potential and practice highlights the need for intentional content adaptation.

Alignment with real-world practice

The project aligns well with broader educational goals, such as improving access to early literacy tools and supporting individualized learning. Interviewees see particular promise for under-resourced classroom settings, where the tool can help bridge gaps in teacher availability or reading time. Dr. Gamiet notes the potential for the tool to slow down the pace of digital content consumption, encouraging children to listen attentively and process stories more deeply, which aligns with current educational best practices. This slower, more intentional form of digital engagement is seen as especially valuable given the fast-scrolling tendencies of most screen-based tools aimed at children.

However, the tool is not yet ready for direct adoption in most real-world contexts. As Dr. Grove points out, the proof of concept meets its technical exploration goals but requires significant repackaging, especially in terms of user experience and accessibility, to be usable by non-technical audiences. Visser also stresses the need for cloud or web deployment and minimizing setup friction for schools or home users, especially in contexts with limited technical infrastructure. Factors like device compatibility, internet access, and setup complexity are repeatedly identified as potential deal-breakers for successful adoption, particularly in rural or low-resource environments.

Alternative approaches and suggestions

All participants offer concrete suggestions for refinement and expansion, including:

- Adding simple visuals or animations to accompany the story text, catering to different learning styles (auditory and visual).
- Including voice-guided navigation for children who cannot yet read, making the tool more accessible to pre-literate users.
- Supporting language selection, including underrepresented languages, to enhance inclusivity and local relevance.
- Building in feedback loops for children (“Did you like this story?”) and adults (teacher or parent review before playback), enabling iterative improvement and safeguarding quality.
- Providing a “teacher mode” for content review and approval before narration could help maintain a healthy balance between AI-driven tools and essential human involvement.
- Considering a hybrid open-source model with optional hosted deployment and premium features to maximize accessibility while supporting sustainability and further development.

Several interviewees also recommend phased testing in classroom environments to gather observational data on usability and learning outcomes, offering a form of validation that extends beyond expert interviews.

These recommendations collectively point toward a more robust, inclusive, and user-friendly version of the tool. Such a version could realistically support both individual learning and classroom integration. They also reflect a shared desire to balance technical innovation with educational integrity, ensuring that the tool serves not only as a showcase for AI but as a practical, empowering resource for learners and educators alike.

Sub-question context

The expert validation process and resulting analysis directly address the following sub-question:

“What results emerge when using the application to support children’s reading skills, and what conclusions can be drawn from these results?”

While the application has not yet been tested with children in a live classroom setting, expert feedback provides valuable insight into its potential educational impact. The narrated stories, age-appropriate vocabulary, and interactive metadata system are recognized as features that align well with practices used to support early literacy. Interviewees note that narrated content can help improve listening comprehension, story awareness, and vocabulary acquisition, especially for young learners who are not yet confident readers. These findings suggest a strong conceptual foundation for supporting reading development, though future classroom trials will be needed to assess measurable outcomes.

4.4 Implementation challenges and practical barriers

While the AI storytelling proof of concept demonstrates strong potential, several critical challenges must be addressed before the system can be deployed in everyday educational or home environments. Interviewees identified a range of technical, usability, and contextual barriers that must be addressed to make the tool both accessible and sustainable.

A major technical concern is the tool's current hardware requirements. Running the application locally requires a relatively powerful machine, especially during story generation, narration, and music synthesis, which rely on large language and audio models. Rudolf Visser emphasizes that most schools and households are unlikely to have access to GPU-equipped devices, making a scalable, cloud-based deployment model essential. In addition, the current models, particularly for music, remain limited. MusicGen-Small, for instance, only produces short clips, and the TTS system (MeloTTS) lacks voice diversity and emotional nuance. These limitations, according to Dr. Grove and Daniels, reduce both the system's immersive quality and its adaptability across varied user contexts.

Usability and accessibility have also been raised as major concerns. The existing Gradio interface, while suitable for prototyping, lacks essential features for children or non-technical users. Interviewees recommend the development of a more intuitive, touch-friendly interface with visual storytelling elements, voice navigation, and tailored onboarding. They also propose creating a separate dashboard for educators or parents to support content oversight and enable different user modes.

The interviews also point out deployment challenges in educational environments, especially in under-resourced areas. Dr. Gamiet stresses that many teachers, particularly in rural schools, may not have the training or infrastructure to adopt new digital tools. As emphasized by multiple interviewees, such tools should enhance rather than replace human involvement in learning environments. Limited internet access, outdated devices, and restrictive school policies further complicate implementation. Proposed solutions include pre-generated content libraries, offline compatible versions, and teacher training materials to reduce adoption friction.

Another key challenge lies in cultural and linguistic adaptability. Without specific prompt engineering, generated stories tend to reflect Western-centric defaults, which may not resonate in diverse classrooms. Both Dr. Gamiet and Daniels advocate for localized metadata and expanded language support, particularly for local underrepresented languages.

Lastly, the project's long-term sustainability depends on its ability to evolve and maintain quality over time. While open-source development encourages accessibility and collaboration, it can also introduce risks related to version control, content moderation, and licensing. Visser and Dr. Grove recommend exploring hybrid models that keep the core system open but monetize additional features or hosting to fund ongoing maintenance. Daniels adds that clear documentation and structured licensing will be essential to support responsible scaling and external adoption.

4.5 Societal and educational value

Despite being in an early prototype phase, the AI storytelling tool already presents meaningful opportunities to address real-world educational and societal challenges. Interviewees consistently emphasized its potential to support early literacy, promote inclusive access to learning materials, and foster creativity among young learners.

Narrated, customizable stories offer a practical solution in classrooms with high student-to-teacher ratios and in homes where adult supervision is limited. Dr. Gamiet and André Daniels observe that such tools can help close literacy gaps by supplementing, but not replacing, human interaction. By enabling children to listen to age-appropriate, engaging content, the system supports vocabulary development, listening comprehension, and story structure awareness.

As noted earlier, the system currently defaults to Western-centric story structures, but its modularity offers opportunities for future adaptation to local languages and culturally resonant content. Dr. Gamiet and Daniels stress the importance of such localization, particularly for learners in multilingual or underrepresented communities. Incorporating African languages and culturally resonant themes could significantly enhance the tool's relevance and accessibility within African educational contexts. Likewise, similar adaptations could be made for other cultural settings, enabling the tool to support localized storytelling approaches and languages across diverse global regions.

Beyond literacy, the customizable structure of the tool is seen as fostering active engagement and creativity. Daniels suggests that features like user-driven plot choices, drawing activities, or alternative story endings could enrich the learning experience and develop critical thinking skills. Visser and Dr. Grove also highlight its potential use in language learning, elder care, and informal education settings, further underscoring its broader societal impact.

However, all experts agree on the importance of responsible and ethical use. They warn against over-reliance on AI for educational engagement and stress the need for adult involvement, content moderation, and clear guidance on safe use. While the system could widen access to educational opportunities, it must be implemented with care to avoid inadvertently reinforcing digital inequities or cultural bias.

In summary, the project demonstrates real potential to support inclusive education, bridge literacy gaps, and broaden access to engaging learning tools. These educational and societal impacts align strongly with the research objectives and point toward valuable directions for future development. Such an approach has the potential to make quality educational materials more accessible, particularly in regions where commercial solutions are unavailable or unaffordable.

Sub-question context

This reflection also contributes to answering the following sub-question:

“How can feedback, both during development and through a user-facing feedback function, be used to improve the AI models and the overall storytelling experience?”

User and expert feedback is already central during the prototype phase, but future versions of the tool could benefit from built-in, structured feedback mechanisms. Interviewees suggest incorporating lightweight feedback options for children, such as simple rating systems or emoji responses, alongside more detailed input from adults. This feedback could be used to refine prompts, improve narration style, and adjust the interface based on real-world usage. Over time, it may even inform fine-tuning of the AI models to better suit user preferences. Embedding feedback loops into the app itself would not only enhance quality and relevance but also encourage continued user trust and engagement.

4.6 Conclusion

This chapter has provided a critical reflection on the development, validation, and real-world potential of the AI storytelling proof of concept. Drawing on feedback from both educational and technical experts, the analysis highlights the project’s core strengths, such as modularity, adaptability, and its potential to support early literacy, as well as key limitations, including hardware requirements, user experience shortcomings, and the need for greater cultural and linguistic inclusivity.

In response to the central research question, the project demonstrates strong potential to contribute to inclusive, scalable literacy tools across diverse educational contexts. While further development and responsible implementation are still needed, the work presents a validated, modular approach to AI-driven storytelling that may serve as a foundation for future educational technology efforts.

The following chapter presents concrete recommendations to address the challenges identified and guide future iterations of the tool. Ultimately, this project illustrates how emerging AI tools, when developed thoughtfully and in consultation with real-world stakeholders, can play a meaningful role in shaping the future of inclusive digital education.

5 Recommendations

5.1 Purpose of this chapter

This chapter translates the insights and lessons learned throughout the development of the AI storytelling tool into a structured set of actionable recommendations. These suggestions are specifically aimed at individuals and organizations interested in developing, implementing, or adopting AI-driven storytelling tools designed for young children. Target audiences include educators, software developers, digital learning specialists, and other practitioners working at the intersection of technology and early education.

They aim to bridge the gap between critical reflection and practical application. They are grounded in the project's internal findings, spanning technical experimentation, system validation, and usability considerations, as well as in the external feedback gathered through expert interviews. By incorporating both perspectives, this chapter presents a credible foundation for improvement.

The primary aims of this chapter are threefold:

1. To offer concrete, research-backed advice for enhancing and deploying AI storytelling tools in real-world environments such as schools, libraries, and homes. These recommendations pay particular attention to factors like literacy development, accessibility, and user engagement.
2. To address key implementation challenges identified during the development process and expert interviews. These include technical limitations, user interface design, scalability, cultural inclusivity, and safeguarding content quality.
3. To support future work in this area by highlighting opportunities for further research, collaborative development, and expanded functionality that could increase the educational and creative value of such tools.

Organized around core themes, educational adaptation, technical and UX improvements, content safety, feedback mechanisms, and future directions, this chapter aims to serve as a practical roadmap. By following these recommendations, interested parties can more effectively harness the potential of AI storytelling to foster literacy and creativity in diverse educational contexts.

5.2 Educational and cultural adaptation

AI-driven storytelling tools offer considerable potential to support early literacy, foster creativity, and promote cultural inclusivity, especially when their design is guided by educational objectives and local relevance. This section outlines concrete strategies for adapting such tools to better serve diverse learners, engage young audiences, and reflect the cultural realities of the communities in which they are used.

Supporting literacy and fostering engagement

To contribute meaningfully to early childhood education, AI-generated stories must reinforce core literacy skills such as vocabulary development, listening comprehension, and narrative awareness. Expert feedback confirmed that using age-appropriate language, clear sentence structures, and a consistent three-act format can significantly support literacy acquisition, particularly when aligned with established frameworks.

Collaboration with educators is essential to ensure classroom relevance. By mapping story themes and vocabulary to grade-level standards, such as those in South Africa's CAPS document, the tool can complement curriculum objectives and developmental milestones. In this way, AI storytelling can move from an experimental novelty to a valuable educational resource. Learning outcomes can be further supported by incorporating light assessment features. Examples include multiple-choice questions or creative prompts where children are invited to retell or draw scenes from the story. These tools not only reinforce comprehension but also help educators monitor progress in an entertaining and engaging way.

Beyond academic alignment, the tool must also encourage active participation. Offering personalization options, such as choosing settings, characters, or themes, can help children feel ownership over the narrative, increasing engagement and reinforcing understanding of story structure. Experts note that interactive storytelling features, like branching narratives where users shape the story's direction, could further enhance motivation and critical thinking. While not yet implemented in the current prototype, such features are frequently highlighted as promising additions.

Although the tool currently centers on audio and text, expanding into multimodal formats could further increase accessibility and engagement. Adding simple visuals, animations, or tactile elements would allow younger or pre-literate children to connect more deeply with the content.

Ensuring inclusivity and cultural relevance

For an AI storytelling tool to be truly inclusive, it must reflect the linguistic and cultural diversity of its users. Expanding language support is a crucial first step. Enabling the generation and narration of stories in local languages can make the tool significantly more accessible in multilingual regions. Though this presents technical challenges, such as the need for compatible multilingual models and TTS systems, the educational and social benefits are substantial.

Cultural relevance goes beyond language. Prompts and metadata should be adapted to reflect local traditions, character types, and environments. Experts emphasized the importance of moving away from default Western narratives in favor of stories that feature familiar elements, such as local animals, holidays, community heroes, or folklore. Such adaptations help ensure the tool feels grounded in children's everyday experiences. To guarantee authenticity and cultural sensitivity, it is advisable to collaborate with local educators, language departments, and cultural organizations. These stakeholders can offer valuable insights during content creation and validation, helping to prevent bias and ensure respectful representation.

Inclusion also involves accessibility for children with different learning needs. The tool should offer flexible settings, including adjustable narration speed, voice guidance, subtitles, and support for assistive technologies such as screen readers or sign language overlays. Allowing educators or caregivers to modify the story's complexity and length ensures content remains appropriately challenging. This flexibility supports engagement across a broad age and ability spectrum.

Summary

By prioritizing curriculum alignment, child engagement, multilingual support, and inclusive design, AI storytelling tools can evolve into powerful resources for early education. These recommendations are grounded in both internal findings and expert feedback, offering a roadmap toward tools that are not only technologically innovative but also educationally meaningful and culturally responsive. While meaningful content is essential, the success of AI storytelling tools also depends heavily on their technical foundation and overall usability, a focus explored in the following section.

5.3 Technical and UX improvements

For AI storytelling tools to progress beyond the prototype phase and deliver lasting value in educational and home settings, they must be supported by a foundation of robust technical infrastructure and thoughtful, user-centered design. This requires not only ensuring system reliability and scalability but also creating an interface that is intuitive, accessible, and engaging for children, educators, and caregivers.

From a technical perspective, cloud-based deployment is highly recommended to meet the demands of real-world use. In environments such as classrooms or public libraries, where multiple children may use the tool concurrently, cloud infrastructure enables seamless scalability. However, in low-resource settings with limited internet access, an offline or hybrid mode should also be considered to ensure universal access. A cloud setup also simplifies ongoing maintenance through automated updates and centralized version control, eliminating the need for local installation or manual configuration. In parallel, adopting a modular architecture strengthens the system's flexibility and maintainability. Separating core components, such as story generation, speech synthesis, and music generation, allows developers to update or replace individual modules without impacting the rest of the pipeline. This approach supports long-term adaptability, making it easier to integrate new features, expand language support, or experiment with improved AI models in the future.

Due to the sensitive nature of working with children's data, privacy and security must be prioritized. All stored and transmitted information should be encrypted, with access to personal or usage data strictly limited to authorized systems or personnel. When possible, anonymizing data helps reduce risk. Compliance with data protection regulations such as South Africa's POPIA or the European Union's GDPR is essential, and transparent communication with caregivers and educators about data collection practices helps build trust and legitimacy.

The user interface should be designed to accommodate the specific needs of young children, particularly those who are pre-literate. Navigation should be simple and linear, supported by large buttons, intuitive icons, and minimal text. Spoken prompts, animations, and visual cues can help guide children through the storytelling process in a way that feels intuitive and rewarding.

To further improve accessibility and engagement, the tool should support multimodal interaction, allowing children to engage via touch, voice, or visual selection, consistently accompanied by high-quality narration. The ability to adjust narration speed or select from different voices and accents not only improves usability but also increases the cultural relatability of the content. Customization options play a vital role in making the tool inclusive for children with varying needs. Educators or caregivers should be able to adjust text size, contrast, narration pace, and language selection. Additionally, features such as subtitles and compatibility with assistive technologies, including screen readers or switch-access devices, can ensure the tool is usable by children with sensory or cognitive challenges.

System feedback and error messages should be friendly, constructive, and reassuring, especially for young users. Instead of displaying technical messages, the tool can use gentle sound effects, animations, or spoken cues to help children recover from mistakes in a reassuring manner. This improves both usability and the child's confidence during interaction. For the tool to be viable in formal educational settings, integration with existing platforms such as Google Classroom or digital library systems should be considered. This not only streamlines adoption for educators but also supports learning continuity by allowing teachers to track usage, manage student progress, and align storytelling activities with lesson objectives.

Long-term sustainability requires a clear plan for ongoing technical support and feature maintenance. This includes thorough documentation, a structured approach to bug tracking, and regular updates based on evolving user needs and feedback. A strong technical foundation paired with a responsive design and update strategy ensures the tool remains functional, relevant, and adaptable to future needs.

By investing in both technical stability and user experience, developers can create AI storytelling tools that are not only innovative but also practical, inclusive and well-suited for real-world deployment in diverse educational contexts. Once a stable technical foundation and child-friendly design are in place, the next priority is to ensure that content is safe, ethically appropriate, and properly validated.

5.4 Content safety, ethics, and validation

With a robust technical foundation in place, the next priority is to ensure that each story is not only engaging but also safe and age-appropriate for young audiences. This section outlines the safeguards, responsibilities, and validation mechanisms necessary to support the responsible use of AI in early childhood storytelling.

Automated safeguards and human oversight

Ensuring that AI-generated stories are age-appropriate, culturally sensitive, and free from harmful content is a critical priority. Both internal testing and expert interviews identified content safety and ethical guardrails as essential for building trust among educators, parents, and learners.

A robust content validation pipeline should operate at multiple points in the story generation process. In the current prototype, this begins with carefully engineered prompts that guide the AI toward using simple, appropriate language and avoiding sensitive or harmful topics. Once a story is generated, automated validation checks are applied, including readability metrics (e.g., Flesch Reading Ease score), grade-level alignment, word count limits, and a comprehensive prohibited words list. This list helps exclude inappropriate language, references to violence, adult themes, or culturally insensitive content, with the list regularly updated to reflect evolving risks. If any check fails, the system automatically regenerates the story, ensuring that only validated outputs reach the user.

However, automation alone may not catch every issue. Experts such as Wouter Grove and Rudolf Visser emphasized the importance of combining algorithmic validation with optional human-in-the-loop review. For example, educators or caregivers could preview stories before they are narrated or flag problematic content to support future refinement. This hybrid model, blending automated filters with selective human review, offers a scalable yet flexible approach to safety. This ensures that both automated and human insights contribute to safeguarding quality, particularly in edge cases or culturally sensitive contexts.

Ethical considerations and transparency

Beyond safety, developers must be attentive to ethical dimensions such as bias, representation, and data handling. AI models can unintentionally reinforce stereotypes or over-represent Western narratives if prompts and metadata are not adapted to reflect local contexts. To address this, story templates and character archetypes should be localized, and cultural validation should be built into the quality assurance process.

Transparency is equally important in building trust. All data collection and processing practices must be clearly communicated to users, particularly adults managing children's access. Privacy protections, including encryption, anonymization, and informed consent, should be standard. Licensing and copyright information for both models and generated content should be clearly displayed in the interface, promoting legal compliance and ethical transparency.

Finally, safety and ethical alignment must remain dynamic. Feedback mechanisms such as user flagging, star ratings, or educator review dashboards can help reveal edge cases and inform adjustments to prompts, filters, or even model retraining over time. This ongoing responsiveness strengthens the system's integrity as it grows in reach and complexity.

Summary

By combining prompt engineering, automated validation, and optional human oversight, AI storytelling tools can consistently deliver content that is safe, inclusive, and age-appropriate. These measures not only protect young users but also build confidence among educators and caregivers, encouraging responsible adoption of AI in early learning environments. With this foundation in place, the next step is to explore how feedback loops can drive continuous improvement across technical, educational, and content dimensions.

5.5 Collecting feedback and iterative improvement

Building an effective AI storytelling tool for children is not a one-time effort but an ongoing process. Continuous feedback and refinement are essential to ensure the system remains educationally relevant, safe, and engaging over time. Both internal reflection and expert interviews emphasized that iterative development, driven by real-world user input, is key to long-term success and adoption.

Designing accessible feedback channels

To keep the tool responsive to its users, feedback mechanisms should be integrated directly into the user interface in ways that accommodate different audiences. For children, lightweight prompts such as star ratings, emoji reactions, or simple questions like "Did you enjoy this story?" can encourage interaction without relying on advanced literacy. These tools allow even young users to communicate enjoyment or confusion, helping identify areas where engagement could be improved.

For adult users, such as teachers, caregivers, or education professionals, more detailed forms or dashboards should be provided. These interfaces enable more nuanced feedback, such as flagging inappropriate content, suggesting feature improvements, or reporting technical issues. This dual feedback structure ensures that input from all key stakeholders supports both day-to-day usability and long-term development. Incorporating multilingual feedback options, especially for adult users in multilingual regions, can also increase participation and help surface region-specific concerns that might otherwise be overlooked.

Turning feedback into action

Feedback should be treated not as a passive data collection step but as a proactive driver of model and UX improvements. Reports of low-quality content, unclear narration, or UI usability issues should be systematically reviewed and acted upon. For example, if educators consistently flag a recurring cultural bias or stereotype in story outputs, this should drive adjustments to prompt templates or even guide future fine-tuning of AI models on culturally balanced datasets. Similarly, repeated feedback on narration clarity or background music transitions could inform updates to the audio synthesis or music module components.

The goal is to establish a feedback-to-update loop where reported issues directly inform model fine-tuning, prompt redesign, validation criteria, or user interface changes. Involving domain experts, such as teachers, language professionals, and cultural consultants, can further enrich this process, especially when introducing new features or adapting the tool for new audiences.

Experts interviewed during this project, including André Daniels and Wouter Grove, stressed that early-stage systems rarely account for all classroom dynamics or student needs. They advocated for a flexible development mindset where real-world usage data and educator insight drive the majority of updates after launch. This ongoing process should also foster transparency and community trust. Informing users about how their feedback has influenced changes, whether through update notes, tooltips, or classroom emails, can encourage continued engagement and signal that their input is valued.

Summary

By embedding accessible feedback channels and committing to iterative, community-informed development, AI storytelling tools can remain safe, inclusive, and educationally meaningful as they evolve. This continuous learning approach not only enables rapid response to emerging issues but also ensures the system adapts to the needs of its users and the broader educational context. As the tool matures, a well-structured feedback and improvement framework will be essential for sustaining its long-term relevance and impact. Ultimately, a robust feedback ecosystem is key to both continuous improvement and lasting user trust.

5.6 Future development and research directions

While the current AI storytelling tool demonstrates strong potential as a proof of concept, both internal evaluation and expert feedback highlight numerous opportunities for further development. The following recommendations aim to extend the tool’s educational value, inclusivity, and long-term sustainability.

Enhancing interactivity, personalization, and accessibility

One major area for future expansion involves adding new modalities and interactive features. Experts consistently recommended visual enhancements, such as illustrations or light animations, to support different learning styles and improve accessibility for pre-literate users. Interactive elements like branching narratives or child-driven plot decisions could further deepen engagement and encourage critical thinking. Additional features such as “karaoke-style” text highlighting, drawing prompts, or personalized naming within stories may help bridge narrative comprehension and early literacy development.

Personalization also plays a key role in making storytelling feel more relevant. Future versions could include user profiles, adaptive difficulty based on reading level or age, and preferences for voice or theme selection. Accessibility improvements, such as sign language overlays, screen reader compatibility, and exploratory options like Braille-friendly outputs, can help extend the tool’s reach to children with a broader range of learning needs.

Supporting linguistic and cultural diversity

To make the tool meaningful in more global and multilingual contexts, broader language support is essential. This includes not only generating and narrating stories in underrepresented languages but also localizing prompts, characters, and story themes to reflect regional cultures. Collaborating with local educators, language experts, and cultural organizations is vital for ensuring that generated content is not only accurate but also respectful and representative.

Additionally, research into the educational impact of multilingual AI-generated content in real classrooms could yield valuable insights. Studies might explore how exposure to stories in a learner’s home language affects engagement, comprehension, or vocabulary acquisition.

Audio, deployment, and long-term sustainability

Improvements to the audio system, including dynamic soundtracks, personalized narration, and ambient sound effects, are frequently cited in expert interviews as key areas for future work. As music generation models become more capable, developers could experiment with longer, context-sensitive background scoring to enrich emotional tone and immersion. Voice cloning, especially for family members or educators, could offer a promising path to creating a more familiar and emotionally engaging listening experience.

For practical deployment, cloud-based delivery remains the most scalable option, especially for use in classrooms or resource-constrained environments. Offering pre-generated content libraries, offline-compatible versions, and detailed teacher training materials would ease adoption and integration. Sustainable development may benefit from a carefully balanced hybrid model, combining the flexibility of open-source contributions with the stability of commercial support.

To guide this evolution, further research is needed, not only on pedagogical outcomes such as literacy gains or creativity enhancement, but also on ethical considerations such as bias mitigation, content validation, and long-term trust in AI-driven learning tools.

Summary

By pursuing these development and research directions, AI storytelling tools can become more inclusive, immersive, and adaptable to diverse educational settings. Continued collaboration with educators, cultural partners, and children themselves will be key to ensuring the tool grows in a way that remains both technologically relevant and pedagogically grounded. Taken together, these directions set the stage for the continued evolution of AI storytelling, one that the final chapter will briefly reflect on in conclusion.

6 Conclusion

6.1 Restating the research question

This thesis set out to explore how AI-driven storytelling tools can be designed and implemented to support early literacy and foster creative engagement among young children. The central research question guiding this project is:

"How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?"

To address this question, the project combined hands-on technical experimentation with insights from educational theory and expert consultation. Its primary objectives are fourfold:

- To develop a functional proof of concept capable of generating narrated stories with background music.
- To assess the educational and technical feasibility of such a tool.
- To ensure content safety and ethical compliance, particularly for young audiences.
- To generate actionable recommendations for future development and implementation.

At the heart of this work is the recognition of a growing need for adaptable, culturally responsive digital learning tools, particularly those capable of enriching classroom and home environments with inclusive, age-appropriate storytelling experiences.

6.2 Summary of key findings

The development and evaluation of the AI storytelling tool revealed key insights across technical, educational, and user experience domains. First, the project confirmed that it is technically feasible to generate and narrate age-appropriate stories with synchronized music transitions through a modular AI pipeline. Built-in validation processes, including readability scoring, grade-level analysis, and content filtering, proved effective in ensuring that outputs remained safe, simple, and appropriate for young audiences.

From an educational standpoint, expert feedback confirmed the tool's potential to support early literacy and encourage creative participation. The structured three-act format, use of simplified language, and story guidance options were all identified as beneficial for engaging children aged 5–8. In particular, the ability to localize stories by language, culture, or theme through metadata adjustments emerged as a key strength for promoting relevance and inclusivity.

User experience considerations highlighted the importance of intuitive, child-centered design. Experts emphasized the value of features such as large interface buttons, spoken prompts, and multimodal interaction pathways, particularly for enhancing accessibility and engagement in future implementations. Accessibility options, such as adjustable narration speed, subtitles, and voice variety, further broadened usability across diverse learning needs. At the same time, the evaluation revealed areas requiring further refinement, including improving coverage for edge cases in content validation, enhancing multilingual functionality, and streamlining navigation for young users.

Finally, the external validation emphasized the importance of continuous feedback and iteration. Integrating accessible feedback channels for both children and adult users was seen as essential for maintaining educational quality, improving safety measures, and adapting the tool over time based on real-world usage.

6.3 Answering the research question

The results of this project demonstrate that AI can effectively generate and narrate children's stories with synchronized background music, provided that careful attention is given to both technical design and educational context. The modular pipeline developed for this thesis enabled the automated creation of age-appropriate, engaging narratives while integrated validation steps ensured content safety and developmental suitability. These outcomes confirm the technical feasibility and practical potential of such tools in both educational and home environments.

From an educational standpoint, the tool's structured story format, simplified language, and personalization features are validated by expert feedback as beneficial for literacy development and creative engagement among children aged 5–8. The ability to adapt content to different languages and cultural settings further enhanced its relevance and inclusivity, addressing a clear need in diverse classrooms and learning contexts.

As a result, the project successfully achieved its four primary objectives: developing a functional proof of concept, assessing its feasibility, ensuring safety and ethical compliance, and generating actionable recommendations for future development. Nonetheless, the project also revealed several limitations. While the automated validation system performed well overall, it occasionally missed subtler cultural or contextual concerns, highlighting the value of optional human oversight. Similarly, while support for multilingual content and accessibility features showed promise, these components require further development to fully meet the needs of all learners. These limitations underscore the importance of iterative refinement and sustained stakeholder collaboration.

Taken together, the findings show that an AI storytelling system, when thoughtfully designed, validated, and continually improved, can generate safe, inclusive, and educationally valuable stories for young children. Together, the project’s strengths and limitations offer a practical foundation and a clear roadmap for the responsible adoption and continued evolution of such tools in real-world educational settings.

6.4 Broader implications and limitations

This project contributes to the expanding field of AI-driven educational technology by demonstrating how generative models can be applied to create engaging, age-appropriate storytelling experiences for young children. By combining technical innovation with educational theory and ethical safeguards, the tool offers a proof of concept for digital resources that are not only interactive and adaptive but also culturally and linguistically responsive. The framework and development approach outlined in this thesis point toward broader applications beyond early literacy, including personalized learning, creative expression, and inclusive content design in a variety of educational settings.

Nonetheless, several limitations must be acknowledged. The tool’s evaluation relied primarily on expert feedback and personal developer testing rather than classroom trials or longitudinal studies. The testing done includes structured feedback from expert interviews and demo sessions, which offered initial insights into usability and interface design. As a result, some findings may not fully reflect the complexities and variability of real-world educational environments. In addition, technical constraints, such as the current limitations of AI-generated music and the scope of multilingual functionality, currently restrict the tool’s adaptability and reach. While the built-in validation system proved effective for basic content filtering, it may still overlook more subtle cultural or contextual issues, underscoring the ongoing need for human oversight and continuous refinement.

Despite these constraints, the project establishes a meaningful foundation for further research and development. It illustrates both the potential and the challenges of applying AI in early childhood education and emphasizes the importance of sustained evaluation, stakeholder feedback, and interdisciplinary collaboration in building responsible, inclusive, and impactful educational technologies.

6.5 Integration of recommendations

The findings and reflections presented in this thesis directly informed a set of actionable recommendations to guide the future development, deployment, and evaluation of AI storytelling tools for children. Central among these is the need to maintain a robust, modular technical architecture that supports scalability, cloud-based deployment, and ongoing updates. Such an infrastructure enables the tool to adapt flexibly to evolving educational needs, technological progress, and user expectations.

Equally important is the prioritization of child-centered user experience and accessibility. Key recommendations include designing intuitive interfaces with large, clearly labeled controls; supporting multimodal interaction (e.g., touch, voice, and audio feedback); and offering customization options to accommodate diverse learning needs and preferences. Expanding language support and improving cultural relevance through localized prompts, inclusive metadata, and collaboration with local communities are also identified as essential for equitable and meaningful use.

To ensure content safety and ethical integrity, the thesis proposes a dual-layer validation system that combines automated checks with optional human oversight. This hybrid model helps address the limitations of algorithmic validation, particularly in catching nuanced cultural, contextual, or representational issues. In parallel, embedding accessible feedback mechanisms for both children and adults and using this input to inform regular system updates is emphasized as a key strategy for maintaining educational value, content quality, and user trust over time.

Finally, the recommendations stress the importance of sustained research and interdisciplinary collaboration. Actively involving educators, language experts, and cultural partners throughout the design and evaluation process will be critical for ensuring that future iterations of the tool remain pedagogically effective, culturally sensitive, and ethically sound across diverse learning environments.

6.6 Final reflections and outlook

In summary, this thesis has demonstrated that thoughtfully designed AI storytelling tools can play a meaningful role in supporting early literacy, creative engagement, and inclusive learning for young children. By integrating technical innovation with educational theory, ethical safeguards, and continuous user feedback, the project provides a practical foundation for the responsible and effective use of AI in early childhood education.

Looking ahead, the rapid evolution of AI and digital learning environments will bring both new opportunities and important challenges. The success of these tools will depend not only on technological progress but also on sustained collaboration among educators, developers, and communities to ensure cultural relevance, accessibility, and ethical integrity.

As education systems continue to embrace digital transformation, the lessons and recommendations from this work offer a concrete foundation for building child-centered, adaptable, and trustworthy AI solutions, tools that can meaningfully enrich diverse learning contexts and support equitable access to creative, literacy-focused resources for years to come.

7 References

[1] I. Isozaki, “Understanding the Current State of AI for Story Generation,” Medium. Accessed: Jan. 08, 2025. [Online]. Available: https://isamu-website.medium.com/understanding-ai-for-stories-d0c1cd7b7bdc

[2] Z. Wen *et al.*, “GROVE: A Retrieval-augmented Complex Story Generation Framework with A Forest of Evidence,” Oct. 24, 2023, *arXiv*: arXiv:2310.05388. doi: 10.48550/arXiv.2310.05388.

[3] Z. Patel, K. El-Refai, J. Pei, and T. Li, “SWAG: Storytelling With Action Guidance,” Oct. 07, 2024, *arXiv*: arXiv:2402.03483. doi: 10.48550/arXiv.2402.03483.

[4] M. Ismayilzada, C. Stevenson, and L. van der Plas, “Evaluating Creative Short Story Generation in Humans and Large Language Models,” Nov. 06, 2024, *arXiv*: arXiv:2411.02316. doi: 10.48550/arXiv.2411.02316.

[5] A. Khatun and D. G. Brown, “Assessing Language Models’ Worldview for Fiction Generation,” Aug. 15, 2024, *arXiv*: arXiv:2408.07904. doi: 10.48550/arXiv.2408.07904.

[6] E. Razumovskaia, J. Maynez, A. Louis, M. Lapata, and S. Narayan, “Crosslingual Story Planning and Generation with Large Language Models,” Mar. 25, 2024, *arXiv*: arXiv:2212.10471. doi: 10.48550/arXiv.2212.10471.

[7] Y. Xu *et al.*, “Fantastic Questions and Where to Find Them: FairytaleQA -- An Authentic Dataset for Narrative Comprehension,” Mar. 26, 2022, *arXiv*: arXiv:2203.13947. doi: 10.48550/arXiv.2203.13947.

[8] C. Chhun, F. M. Suchanek, and C. Clavel, “Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation,” May 22, 2024, *arXiv*: arXiv:2405.13769. doi: 10.48550/arXiv.2405.13769.

[9] M. R. Hasanabadi, “An overview of text-to-speech systems and media applications,” Oct. 22, 2023, *arXiv*: arXiv:2310.14301. doi: 10.48550/arXiv.2310.14301.

[10] Y. Ren *et al.*, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” Aug. 08, 2022, *arXiv*: arXiv:2006.04558. doi: 10.48550/arXiv.2006.04558.

[11] Z. Ju *et al.*, “NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models,” Apr. 23, 2024, *arXiv*: arXiv:2403.03100. doi: 10.48550/arXiv.2403.03100.

[12] C. Minixhofer, O. Klejch, and P. Bell, “TTSDS -- Text-to-Speech Distribution Score,” Dec. 02, 2024, *arXiv*: arXiv:2407.12707. doi: 10.48550/arXiv.2407.12707.

[13] J. Ao *et al.*, “SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing,” May 24, 2022, *arXiv*: arXiv:2110.07205. doi: 10.48550/arXiv.2110.07205.

[14] A. Agostinelli *et al.*, “MusicLM: Generating Music From Text,” Jan. 26, 2023, *arXiv*: arXiv:2301.11325. doi: 10.48550/arXiv.2301.11325.

[15] Y.-H. Lan, W.-Y. Hsiao, H.-C. Cheng, and Y.-H. Yang, “MusiConGen: Rhythm and Chord Control for Transformer-Based Text-to-Music Generation,” Jul. 21, 2024, *arXiv*: arXiv:2407.15060. doi: 10.48550/arXiv.2407.15060.

[16] P. Zhu *et al.*, “ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models,” Sep. 21, 2023, *arXiv*: arXiv:2302.04456. doi: 10.48550/arXiv.2302.04456.

[17] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, “Moûsai: Efficient Text-to-Music Diffusion Models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8050–8068. doi: 10.18653/v1/2024.acl-long.437.

[18] K. R. Prajwal *et al.*, “MusicFlow: Cascaded Flow Matching for Text Guided Music Generation,” Oct. 27, 2024, *arXiv*: arXiv:2410.20478. doi: 10.48550/arXiv.2410.20478.

8 Appendices

8.1 Reports guest speakers

Guest lecture report – XR projects for the future

Speaker: Pieter Van Leugenhagen (Yondr)
Date: January 15, 2025, 13:00
Location: Online
Topic: XR Projects for the Future

Introduction

As part of the Research Project module, students attended a series of guest lectures on future-oriented technology. On January 15, 2025, Pieter Van Leugenhagen, co-founder and managing partner at Yondr, delivered a session titled "XR Projects for the Future." Yondr is a Belgian immersive media company specializing in Extended Reality (XR), including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) [1].

This lecture explored the ongoing evolution of XR, moving beyond entertainment into areas such as education, training, and workplace collaboration. Pieter discussed real-world use cases and addressed both the benefits and limitations of XR adoption, particularly in professional and academic settings. The session was highly relevant for students interested in immersive media, spatial computing, and the future of user interaction in 3D environments.

Spatial computing: the next digital paradigm

Pieter introduced spatial computing as the next major evolution in digital interaction. Spatial computing is a 3D-centered approach that uses artificial intelligence, computer vision, and XR to seamlessly blend virtual experiences into the physical world. This paradigm aims to bridge the gap between the 3D nature of our reality and the predominantly 2D way we currently interact with digital content [2].

He emphasized that, although we live in a 3D world, most digital interactions still happen on flat screens. XR and spatial computing have the potential to make digital interactions more immersive and context-aware, fundamentally changing how we experience and use technology [3].

XR devices and use cases

Today, XR is mostly accessed through headsets, which are used for entertainment, productivity, and communication. However, Pieter highlighted the emergence of lightweight smartglasses, such as Ray-Ban smart sunglasses, which are already capable of making calls, taking photos and videos, and playing music. Some of these devices are beginning to integrate generative AI features, enhancing their utility for hands-free tasks and real-time information access [4].

Pieter also pointed out that XR is not limited to dedicated headsets; many people already use XR-like features on their smartphones, such as AR filters and navigation apps. This demonstrates that spatial computing is becoming increasingly mainstream, even if not always in fully immersive forms [5].

Key concepts and trends

- Phygital experiences: Spatial computing enables blended digital-physical experiences, often referred to as "phygital." This concept is already being realized in virtual showrooms and interactive retail environments [6].
- Platform evolution: Pieter discussed the progression of computing platforms: desktop → laptop → mobile phone → smartglasses. XR and spatial computing represent the next step in this evolution, offering new ways to interact with digital content [6].
- Everyday adoption: Many users are already engaging with XR-like features without realizing it, indicating a gradual but steady adoption of these technologies [3].

Applications and observations

Pieter provided examples of XR in action, including:

- Education: XR can transform learning by enabling immersive simulations and hands-on experiences, making abstract concepts more tangible and engaging [2].
- Workplace collaboration: XR facilitates remote assistance, design prototyping, and digital storytelling, enhancing productivity and creativity [5].
- Accessibility: The potential of XR to improve accessibility and interaction design was also discussed, with the possibility of making digital environments more inclusive [7].

He also noted that several companies have already launched virtual showrooms and other phygital experiences, demonstrating the commercial viability of these technologies [6].

Conclusion & critical reflection

As a student working on an AI-driven children's storytelling system, I found strong parallels between XR and AI storytelling. Both aim to create immersive, emotionally engaging experiences, key components of effective learning tools [4].

This lecture encouraged me to:

- Explore how spatial awareness and presence can enhance my project, for example through audio spatialization or visual immersion techniques [2].
- Consider multimodal interaction, combining voice, visuals, and movement in educational tools [7].
- Stay informed about developments at the intersection of XR and AI, especially as wearable devices begin to support generative AI natively [4].

Overall, the session broadened my perspective on XR, not just as a technical innovation, but as a catalyst for rethinking education, communication, and everyday digital habits. The reminder that we still interact with digital tools in a 2D manner, despite living in a 3D world, was particularly thought-provoking and relevant for future projects [3].

References

[1] "Yondr agency • Immersive & Digital Experiences." Accessed: Apr. 25, 2025. [Online]. Available: https://yondr.agency

[2] "How Spatial Computing and Visual AI Will Transform K-12 Learning | LinkedIn." Accessed: Apr. 25, 2025. [Online]. Available: https://www.linkedin.com/pulse/how-spatial-computing-visual-ai-transform-k-12-learning-mateer-a9e0e/

[3] "What is XR, and how is it radically transforming industries?" Accessed: Apr. 25, 2025. [Online]. Available: https://www.autodesk.com/design-make/articles/what-is-xr

[4] "The Future of XR in the AI Era: A Digital Creator's Perspective | LinkedIn." Accessed: Apr. 25, 2025. [Online]. Available: https://www.linkedin.com/pulse/future-xr-ai-era-digital-creators-perspective-shereen-el-bowety-9gmcf/

[5] "Extended Reality (XR) in Workplace Collaboration: Transforming How We Work Together - GHRC LLP." Accessed: Apr. 25, 2025. [Online]. Available: https://globalhrcommunity.com/extended-reality-xr-in-workplace-collaboration-transforming-how-we-work-together/

[6] M. Ahlström, "Research project focuses on the future of XR and AI," BTH. Accessed: Apr. 25, 2025. [Online]. Available: https://www.bth.se/eng/news/research-project-focuses-on-the-future-of-xr-and-ai/

[7] "The Future of Education: Spatial and VR Computing | LinkedIn." Accessed: Apr. 25, 2025. [Online]. Available: https://www.linkedin.com/pulse/future-education-spatial-vr-computing-jason-benitez/

Guest lecture report – foundations of generative AI

Speaker: Dr. Johan Breytenbach (University of the Western Cape)
Date: April 24, 2025, 13:00
Location: University of the Western Cape (UWC), online
Lecture Title: Foundations of Generative AI

Introduction

This guest lecture, delivered by Dr. Johan Breytenbach at the University of the Western Cape (UWC), provided an academic introduction to generative AI (GenAI). The session was attended as an alternative to a cancelled Howest lecture and was recommended by my internship mentor. It also marked the start of UWC's short course "The Future is GenAI," reflecting the university's commitment to digital education and local expertise in AI.

What is Generative AI?

Dr. Breytenbach began by defining generative AI as artificial intelligence systems designed to create new content, such as text, audio, images, or code, based on input data. Unlike traditional AI, which focuses on classification or detection, GenAI generates original outputs in response to prompts [1], [2].

Generative AI vs. Traditional AI

The lecture highlighted the distinction between traditional AI and GenAI:

- Traditional AI: Used for tasks like classification, detection, and prediction.
- Generative AI: Produces human-like content, such as stories, responses, or images, expanding the scope of what AI can do [1], [2].

Large Language Models (LLMs)

A significant portion of the session focused on large language models (LLMs), which are foundational to text-based GenAI. LLMs, such as GPT and Llama, are built on transformer architectures that use attention mechanisms to understand relationships between tokens (units of text) and context [3]. These models are trained on vast datasets, often containing hundreds of billions of tokens, and require immense computational resources [1], [4].

Key Concepts Explained

- Tokens & Tokenization: LLMs process text by breaking it into tokens, which can be words or subwords. Tokenization is a core step in natural language processing [3].
- Training Process: Training LLMs requires massive, diverse datasets and significant computational power. For example, models like GPT-3 are trained on hundreds of billions of tokens, using both public and synthetic data [1], [4].
- Learning Types:
  - *Supervised learning*: Learning from labeled data.
  - *Self-supervised learning*: Predicting parts of input from other parts, common in LLMs.
  - *Reinforcement learning*: Learning from feedback, such as reinforcement learning from human feedback (RLHF) [2], [4].

Model architectures & challenges

Dr. Breytenbach explained that most GenAI models use neural networks with attention mechanisms (transformers), which allow them to maintain context and generate coherent output [4]. However, challenges include:

- Maintaining long-term context in generated content [3].
- Handling bias or missing context, which can lead to inaccurate or inappropriate outputs [5], [6].
- Balancing objective and stylistic content generation.

Training requirements

To train a GenAI model effectively, several factors are essential:

- Use of tokenized, high-quality datasets [1], [4].
- Computational efficiency and access to powerful hardware (e.g., GPUs) [1], [6].
- Ethical considerations, such as data privacy, bias mitigation, and resource consumption [5], [6].

Conclusion & critical reflection

This session provided a clear, accessible introduction to the foundational concepts of generative AI and LLMs. I gained a better understanding of the terminology, architectures, and challenges involved in building and deploying GenAI systems. The emphasis on ethical considerations and the computational demands of training these models was particularly relevant, given the rapid growth of AI in both academia and industry [5], [6].

As a student focused on AI-driven applications, this lecture encouraged me to:

- Deepen my understanding of LLM architectures and their training requirements [3], [4].
- Consider the ethical implications of AI development, especially regarding bias and data privacy [5].
- Stay informed about advances in GenAI, including prompt engineering and model evaluation, which will be covered in future sessions.

Overall, this lecture provided a strong foundation for further exploration of generative AI and its applications in both research and real-world settings.

References

[1] “What is Generative AI? | IBM.” Accessed: Apr. 25, 2025. [Online]. Available: https://www.ibm.com/think/topics/generative-ai

[2] “What is GenAI? Generative AI Explained | TechTarget,” Search Enterprise AI. Accessed: Apr. 25, 2025. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/generative-AI

[3] “AI Demystified: Introduction to large language models | University IT.” Accessed: Apr. 25, 2025. [Online]. Available: https://uit.stanford.edu/service/techtraining/ai-demystified/llm

[4] “What is LLM? - Large Language Models Explained - AWS,” Amazon Web Services, Inc. Accessed: Apr. 25, 2025. [Online]. Available: https://aws.amazon.com/what-is/large-language-model/

[5] “Understanding the Ethics of Generative AI: Risks, Concerns, and Best Practices.” Accessed: Apr. 25, 2025. [Online]. Available: https://www.datacamp.com/tutorial/ethics-in-generative-ai

[6] “The Complete Guide to Generative AI Architecture.” Accessed: Apr. 25, 2025. [Online]. Available: https://bombaysoftwares.com/blog/generative-ai-architecture

8.2 Interview questions

Introduction & context

Project Clarity: Did you have a chance to review the project materials (video, code repository, project overview)? Is there anything you'd like me to clarify regarding the scope or objectives?

Initial Impressions: Based on what you've seen so far, does the core concept of this AI storytelling application (story generation, TTS, and music) make sense in an educational or industry context? What stood out to you most?

Technical aspects

Pipeline Strengths & Refinements: From a technical perspective, which parts of the pipeline, story generation, text-to-speech, or music, do you see as the strongest? Where do you think there's room for improvement or optimization?

Scalability & Deployment Contexts: What technical considerations come to mind when thinking about deploying this application at scale (e.g., cloud environments) or in low-resource or offline settings? How adaptable do you find the current architecture?

Modularity & Future Adaptability: Do you feel the modular setup supports the integration of newer AI models or technologies over time? And in that context, do you see value in combining modular AI components with human-curated elements?

Educational value & user engagement

Early Literacy Support: Do you believe this application can meaningfully support early literacy, listening comprehension, or imaginative engagement for children aged 5–8?

Creative Structure: How effective do you find the current selection mechanism (setting, characters, theme) in encouraging creativity while staying developmentally appropriate?

Feature Enhancements: Are there any features or changes you'd suggest to make the app more engaging or educationally effective? For example: voice guidance, larger icons, or interactive elements?

Classroom Usability: Do you see this tool aligning well with curriculum goals or literacy frameworks? Would it be practical for teachers to use in classrooms without extensive training?

Business & implementation

Market Fit: How do you see an AI-powered storytelling tool like this fitting into the educational tech or content creation market?

Implementation Challenges: What barriers might schools, publishers, or platforms face in adopting or integrating this system?

Open-Source vs. Commercial: Do you think an open-source version could appeal to institutions, would a hybrid or licensed model be more suitable for broader adoption?

Maintenance & Updates: What do you anticipate in terms of long-term maintenance, model updates, moderation filters, feature requests, and how might that impact adoption?

AI vs. Simpler Solutions: From a business perspective, does the investment in generative AI justify itself compared to simpler template-based tools? Or is the AI element a strong differentiator?

Ethical considerations & validation

Ethical Risks: What ethical concerns do you foresee with AI-generated children's content (e.g., bias, cultural representation, content safety)?

Content Validation Approaches: What are the best practices to ensure safety and quality in this context? Is an automated validation system (e.g., keyword filtering) enough, or should human moderation be included?

Validation & Quality Checks: I currently use readability metrics and audio clarity validation, are these sufficient? What other strategies would you suggest to assess content quality and suitability for children?

User Feedback Loop: What’s the best way to incorporate feedback from end users (educators, parents, children) to continuously improve or retrain the AI models?

Future research and expansion

Emerging Technologies: Where do you see AI storytelling heading next, real-time interactivity, personalized narration, or multimodal formats like AR/VR?

Continuous Background Music: If future models allow for longer, more coherent background music, do you think this would significantly enhance the storytelling experience?

Broader Use Cases: Beyond the current target group, do you see potential for other user groups, like older children, language learners, or special education settings?

Multilingual Storytelling: Do you see market demand or educational value in generating stories in multiple languages or dialects?

Closing reflections

Strengths & Weaknesses: From an overall perspective, what would you consider the strongest and weakest aspects of this project, technically, educationally, or in terms of user experience?

Open Floor: Is there anything else you’d like to add or suggest that we haven’t discussed yet?

Reference Preference: How would you prefer I refer to you in my thesis (name, role, organization)?

Follow-Up Permission: Would you be open to a follow-up conversation if additional questions arise later in the process?

8.3 Interview transcripts

Interview transcript – Gassant Gamiet

Interview with Doctor Gassant Gamiet, part of the faculty of education at UWC, also the president of the coding & robotics club.

LM: I will say, this is the first interview I've done for the thesis, so I don't have much experience with it. I just prepared some questions.
GG: Okay.
LM: Let me start by asking: did you receive the info packet and everything about the project? And did you understand what the project is about, or do you have any questions about it? I have it open here too. The info packet included examples of generated stories and audio, the code itself, an installation guide, a user manual I made, and a short demo video. Everything was clear? You understand what the application is for and how it relates to the research question?
GG: Could you maybe read the research question again, just to refresh my memory?
LM: Sure. The main research question was: "*How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?*". Although, as I mentioned, the background music didn't fully work out due to time constraints, available resources, and model limitations. So, I ended up using short transition music clips instead.
GG: Okay.
LM: Next question, do you think the core concept of an AI story application makes sense in an educational or industry context? Also, was there anything in particular that stood out to you?
GG: To be honest, I didn't go through everything. I just looked at some snippets of it. I think I'd need more time to go through it properly and form an opinion.
LM: Yeah, no worries. If you want, we can quickly go over the demo video together right now.
GG: I'd prefer to browse through the full thing before giving input. But okay, can you show it?
LM: Sure! So, this is the demo video.

[Demo started]

LM: Yeah, all of this is running locally on my laptop, which is a bit older, so it takes a while.
GG: I see, and you're using an Intel i5?
LM: Yes, and it has a GeForce 2060 GPU in it.
GG: Okay. So, you're combining three models?
LM: Yes, Llama for story generation, MeloTTS which is a free open-source model for narration, and Facebook's MusicGen-Small version for the music.
GG: Does the music play in the background of the story?
LM: Not exactly. Like I said, full background music wasn't feasible. The narrated stories are five to seven minutes long, and MusicGen-Small only produces three-second clips. Larger models weren't really an option either due to storage and performance limitations. So instead, I added short music transitions between sections of the story.
GG: Ah, okay. Can I ask questions about this?
LM: Of course, go ahead!
GG: Did you create a storyboard for this? Like a script or visual outline of the story, from beginning to end?
LM: Not exactly a storyboard. I can show you what the prompts looked like. They were structured prompts that indicated things like: "This is the beginning of the story, introduce the characters, set up the setting, and maybe mention a challenge or antagonist". But the story itself was generated by the model based on those instructions.
GG: So, you didn't ask the AI to first create an outline, it just creates it directly?
LM: Yes, I prompted it to generate the story directly. For example, this is the start of one of the prompts: "You are a skilled storyteller writing the beginning of a children's story. Use language suitable for children aged 5–8". Then I include the chosen setting, characters, and theme.
GG: That sounds similar to a storyboard. In film or animation, the storyboard is the timeline they use to map everything out, from the first minute to the last.
LM: Yeah, this was more of a structured prompting method than an actual storyboard, but it does follow a similar logic.
GG: So those six elements you mentioned.
LM: Right, for the beginning prompt, I include six key elements. Then for the middle section, I add the beginning's summary and define 6 new instructions like "introduce a challenge" or "highlight the story theme". The ending section includes a recap of the first two parts, then 5 instructions such as "solve the problem, show

how the characters felt," and always "end with a happy resolution", since it's a story for kids, after all.
GG: Okay, that makes sense. Can I see the rest?
LM: Sure!

[Demo continues]

LM: So yeah, for the Llama model, I used Ollama to call it locally. MeloTTS was a Dockerized image with a REST API, someone else made it, and I just hooked it in. The MusicGen model was used directly via Hugging Face's Transformers package. I have some automated validation of each step as well.
GG: Okay, so your main objective was to create an audio story?
LM: Yes, an audio story. But it also includes the visual text of the story as well. As I mentioned earlier, this is more of a demo or proof of concept. If I were to imagine this as a complete application, I'd mainly have large icons instead of just the emojis I'm currently using, something kids could easily select to create their story. Possibly, I'd also integrate TTS for a full voice-driven experience. Like, "Hi! Welcome to this little storyteller. If you choose some characters here and a theme there, we can create a story for you.". The idea is that the app narrates the story but also shows the text in sync, kind of like karaoke. The text would be highlighted as the words are read, helping children follow along, maybe even start recognizing some simpler words. So yes, narration is the main focus, but I did consider the visual side too.
GG: Right, and the whole process is educational, where children learn through it. Does it have any navigation? Like, can they pause, rewind, or slow down the narration? Like you would with videos on YouTube?
LM: I can show you that quickly. In the front-end, there's a simple audio player, just the default one that comes with Gradio. It includes the full story text in a textbox below the audio. You can pause or change playback speed but I also configured the TTS model to play slightly slower by default, around 0.8 or 0.9 speed, to make it more understandable for the age group.
GG: And your target audience is that 5–8 age category?
LM: Exactly.
GG: And which language is the content in?
LM: It's fully in English for now. My original idea was to make it multilingual, something like *Dora the Explorer*, using simple vocabulary words in a second language. But combining both a multilingual text generation model and a TTS model that supports the same languages wasn't feasible with my resources. So, I settled on a monolingual version. Still, I think it could be adapted if the right TTS model is found or adapted. For example, the AI could replace basic vocabulary words with translations, like if the story mentions "red," and we wanted some Japanese included, it could say "red (aka)" so kids see both terms used equally.
GG: Okay, good. Are you done explaining the process?
LM: Yes, that's pretty much the whole pipeline, it goes through the generation, validation, and finally creates the story and audio output. I can play an example now if you'd like.
GG: Yes, please.
LM: This one is the story used in the demo. It's based on the "Small Kingdom" setting.

[audio of a narrated story]

LM: So yeah, that's an example of one of the generated stories. The TTS quality isn't perfect, sometimes it mispronounces words or stresses the wrong syllables. For example, it might misread "lives" as "lives" or vice versa. But overall, it's decent for an open-source option.

[audio of a narrated story continues]

LM: So that's an example of one the stories.
GG: I'm quite impressed, really.
LM: I'm glad you like it!
GG: Yes so, you had no co-creator or external input in this?
LM: No, the only input comes from the metadata: the setting, characters, and theme with some description for those. Here the theme was likely either "friendship" or "family." Then I used detailed prompt engineering to guide the story generation.
GG: Okay. Have you considered adding images throughout the story?
LM: That is a possibility, but it wasn't the focus. My goal was to focus on narration and storytelling, not visuals. The inspiration for this project came from my childhood, when I used to sleep over at my grandmother's house, she'd read stories to me. Some had illustrations, but mostly it was just her voice and the story. That memory stuck with me.
GG: Ah, I see. That's your objective. Still, I'd suggest trying two versions, one with visuals and one without. Then

you could do a small study with young learners in two groups to see if the visuals help them understand the story better or grasp its moral.
LM: Yeah, that's a good idea.
GG: About the setting, was it meant to be specifically British, or more generic?
LM: Honestly, it's just labelled "Small Kingdom." If I check the metadata, the descriptions are all generic. For example, a wizard is described as "someone wise with magical powers". The "Small Kingdom" is described as "a busy and happy kingdom with a big castle and small houses. Knights, royals, and villagers all going about their day. Adventures often begin here". So, the idea was for it to feel familiar but not be tied to a specific culture or country. I wanted it to be globally adaptable.
GG: And how long does it take to generate a story once the prompts are input?
LM: About 8 to 10 minutes, mostly because of the story generation. The TTS and music each take about a minute, but generating the story takes the longest. That's partly because I use pretty strict prompts to ensure the output is appropriate for children and clear to understand.
GG: How long did the whole process take overall to get this final product?
LM: The research project lasted about a month, so roughly four weeks from start to finish. It was always meant to be a prototype or proof of concept, not a final product. But I'm happy with what I was able to build.
GG: You mentioned Ollama. Is that a specific app or tool for story generation?
LM: Ollama is a tool that lets you run open-source models locally. You can download and run models like Llama, Phi, or Qwen. It's really handy, just a command-line interface where you type "ollama run" and it launches the model. I used it with Llama 3.1, which is one of the top-performing models on Hugging Face's text generation leaderboard. It also supports new models as well like DeepSeek. But for my purposes, Llama worked best. You can use this in your own command prompt as well and it would work the same as with ChatGPT, you just type something in and the model gives a response back.
GG: Can it re-run outputs if you're not satisfied? Like if you want a better version?
LM: Yes, kind of. In my implementation, I set up automated checks, things like Flesch Reading Ease, grade level, and a list of prohibited words (e.g., no "gun," "knife," etc.). If a generated story fails validation, it gets re-run up to a certain number of times until it meets all the criteria.
GG: Have you tried using multiple AI models, like one generating and another reviewing?
LM: That's something I considered. I read a paper that used an "action discriminator", a second AI model that consistently checks in with the first model and if changes are needed. It's an interesting approach, but it wasn't feasible for me to implement. Training and running two models locally would've been too resource intensive.
GG: And what about generating a storyboard from the output? Like an outline or visual timeline?
LM: That's something I didn't really think about. Ollama doesn't do visuals, just text. I suppose I could try asking the model to generate a simple outline or a flowchart-like summary, but I never explored that. In my research, I focused mainly on AI-based story generation, narration, and music, not so much on visual planning.
GG: The storyboard is just the timeline, thinking through the sequence of scenes. Before AI really became a thing, I used to teach digital storytelling for years with my students. We'd create stories using local images taken on their phones, then use an app, at the time I think it was called Storybook. There was also an old Microsoft one. I can't remember the name now, but I'm sure similar AI-based tools exist nowadays.
LM: Yeah.
GG: Photo Story, that was it. A German product, I believe. You'd take still images, and it would animate them, adding pan and zoom effects. You could put the text at the bottom, and now, with your project, the narration could be generated from that text. This whole thing is fascinating. I'd suggest asking the AI apps to help generate those visual plans too, just to better understand the thinking process behind how these stories are constructed.
LM: Yeah, the AI's "thinking" process, so to speak.
GG: Exactly. Since it's a language model, it pulls from all this learned data. You prompted it with certain words, so it pieces those together. Still, you said the application takes a month to develop, but the result is just about four minutes?
LM: Well, that specific story is around five and a half minutes, actually. I've got several in total.
GG: But that was just one example?
LM: Yes. For instance, the underwater one has a mermaid and an octopus inventor. That one's also around five minutes. I have four stories in total, one runs over six and a half minutes, and another, like the flower meadow one, is just under five. The durations vary depending on the word count. I generally target between 800 and 1300 words, 800 being the minimum I allow. I'm not too strict as long as it meets that threshold. Also, the code is open-source. It's all on my GitHub. Anyone interested can try it out.
GG: That's great.

LM: Yeah, the code includes the main app, the metadata for prompts and characters, the front-end, and license information. MeloTTS is under a very permissive license, but MusicGen and Llama are a bit more restrictive. For example, Llama's license requires you to attribute the model and follow usage rules. I made sure to include all of that in the documentation.
GG: Before you started, when doing your research, I assume you looked into similar tools? Other story generators?
LM: Yes, I did look into various apps that generate stories or include videos. But I didn't find many that combined all three elements: story, narration, and music. Especially with narration being an important part of my idea. This research was done back in December and Januari, of course so that might have changed. The use case I envisioned was for parents who want to generate bedtime stories, simple and quick. With TTS voice cloning, it could even sound like the parent. I tested a demo that only needed a 15-second voice clip to clone your voice, and it worked surprisingly well.
GG: That's also a bit scary.
LM: Yeah, definitely.
GG: We were talking about this just yesterday in an honors class I'm supervising. Three students are doing AI research projects, and we discussed how quickly this tech is evolving. You can't always tell if the voice you're hearing is real.
LM: Exactly.
GG: We already know images can be faked. Now voices and even full videos can be AI-generated.
LM: Yeah, and it comes with serious risks. Someone's voice could be cloned to make them say offensive things. Same with fake images, especially those manipulated to create harmful content. It's why ethics and misuse are important to address.
GG: Even voice notes on WhatsApp, you can't always trust they're genuine.
LM: Right. That's why I have a list of prohibited words in my app. No adult themes, horror, hate speech, or prejudice. I originally had "hurt" in the list too but removed it because sometimes the characters get hurt in an innocent way, like scraping a knee. That's still child appropriate.
GG: Makes sense. Can you play the underwater one?
LM: Sure, I'll play a minute or two of that one. I can also show this list of instruments based on the setting which is what is used for the music transition. So underwater would use bubble sounds as one of them.

[audio of a narrated story]

LM: This one actually says, "to be continued," after the beginning output, which it shouldn't do.

[audio of a narrated story continues]

LM: So, for example, it repeats the phrase "flowing hair like a golden river" twice. But overall, the story's pretty solid. Occasionally there are minor flaws like that, or elements that don't fully align. It's not perfect, but quite decent.
GG: Still, very impressive. So, the core of the story was about putting together a music concert?
LM: Something like that. It started off with that intention, they build instruments and talk about performing, but then it shifts. They find a treasure map and get caught up in an adventure. The logic isn't always perfect.
GG: But that's typical for kids' stories. Short attention spans. One minute it's music, the next it's treasure hunting.
LM: Exactly.
GG: Have you tried regenerating the same prompt? Would the story change?
LM: Yes, I have. It does change with each run, but some elements tend to reappear. For example, the female characters are often named Luna, even when I rerun the prompt or change it.
GG: It reminds me of characters like Dory from *Finding Nemo*. It picks up on patterns it's seen in other stories.
LM: Yeah. The front-end has metadata with character descriptions. Like for the underwater setting, it says things like: "a beautiful world under the sea, full of colorful fish…"
GG: But I noticed the bubble sound didn't actually come through.
LM: Yeah, sometimes the music isn't very noticeable. MusicGen only generates 3-second clips, and not all are great. The music was the last part I integrated, and I wasn't 100% sure how to best use it yet.
GG: With audio storytelling, background music can create very different moods. Creepy, happy, mysterious, it all depends on the score.
LM: Yeah, when I first considered integrating background music, I realized I'd have to figure out the mood of each scene. So, I thought about splitting the story into segments and analyzing them separately to determine what kind of music should play in each part.

GG: Can these AI tools actually create those scene-based segments for you?
LM: Definitely. Once the full story is generated, we already have the text. You could use that as a guide.
GG: So, you'd convert that to an audio file and give it to the AI to make edits?
LM: Sort of. It's better if you use the text directly. You'd need to link that with the audio timestamps. Initially, I thought about using something like OpenAI's Whisper, it can transcribe audio and align it with timestamps. So, for example, at 1 minute 45 seconds, we'd know we're entering a new scene and could cue a music change there. That was the idea, until I realized full background music wouldn't really be feasible.
GG: Right. But just wait a day or two, and some new tool will show up that does it perfectly! It's amazing how fast things move. So now that you have all these different stories, what are you planning to do with them? Are they going to be compiled into something or just kept as separate stories?
LM: If I had a fully developed application, I'd definitely have it save each story. Each generated story could be saved with a fitting title, either generated automatically or entered by the user, along with the date it was created. If users had accounts or profiles, like a child's first name, it could say, "This story was made by Timmy on February 25th," for example.
GG: That's a great idea.
LM: And if they really liked a story, they could come back to it. In a full app, you could even request a sequel or reuse the same characters for a new adventure. Maybe even let the user change up parts, like the characters, the setting, or the plot.
GG: Exactly. The user could personalize it more. That would be great in a full version.
LM: Yeah. And I'd also include a simple feedback system, maybe just star ratings. Like, "I loved this story," or "This one wasn't as fun."
GG: Where do you see this kind of tool being used? You said the initial idea came from bedtime stories with grandparents or parents reading aloud. But what about in preschools or kindergartens?
LM: I think it could be adapted in many directions. For home use, it could offer easy entertainment or educational value. For schools or kindergartens, it could be used during reading time, especially for early literacy exposure.
GG: And what about adapting it to different cultures? For example, African stories. Could the flavor of the story change depending on cultural context? Such as recreating forgotten folk tales and such?
LM: Absolutely. If you have those stories somewhere and get them in a proper data format, you could fine-tune the text generation model using those stories. Or you could simply adjust the metadata, so instead of "wizard," you insert a traditional African character or mythical creature, with a short description of its traits. Even within the prompt itself, if you specify that the story should reflect African culture and the model has enough training data, it could generate something culturally relevant and respectful.
GG: That's why I brought this up. We've got 11 official languages in South Africa. English is dominant, but each province has its own major language. Years ago, when Android apps were becoming popular, most of the available content was only in English, Afrikaans, or Zulu, Zulu being the second-most spoken language. There's such a need for stories in other local languages, like Xhosa, especially in the Eastern Cape. There are apps now with Zulu story content, but there's still so much untapped potential. With AI, we could really expand that.
LM: Yeah, I know. The key thing is having enough data, clear, high-quality audio from people speaking languages like Zulu, with proper pronunciation and natural flow. Once you have that, you can fine-tune an existing TTS model on that language. Fine-tuning is far easier than building a text-to-speech model from scratch.
GG: Have you thought about meeting up with the language education department, especially the part focused on African languages. We've got a center specifically for that. It's called the CALT center. We've got two such centers just downstairs.
LM: I'd definitely be interested in meeting with them.
GG: The center I'm thinking of is possibly the largest in South Africa among higher education institutions. Dr. Harun has been running it for over 20 years. Right next to it is the CALT center, Center for African Language Teaching. That one focuses specifically on Xhosa. They publish books, develop early education materials. Dr. Timakozi is very involved in the foundation phase with storytelling. If she hears about your project, I'm pretty sure she'll be interested.
LM: I'd absolutely be open to meeting with her to discuss it. As I mentioned, my university also has its own AI lab, so sharing my work or even collaborating wouldn't be difficult. If someone has a solid base in AI and programming, they could build something similar. I think there's great potential for developing African language storytelling AIs through collaboration.
GG: Yes, there's a big market for it. Actually, I've worked with Dr. Timakozi on coding and robotics in early childhood education. That's why they're heading to Denmark soon, she's presenting at a digital literacy conference in Copenhagen in April. How long are you still here again?

LM: I'll be in South Africa until the end of June. Technically, my internship ends the first week of June. After that I have my final thesis defense and internship presentation. Then I'm heading back June 22nd.
GG: Perfect. If you could meet with her before then, that would be great. Is there anything else you wanted to ask me?
LM: Just a few quick questions, mostly focused on the educational side of things. You've already seen how everything works and what the end product looks like.
GG: Yes, and I'm glad you explained the behind-the-scenes process. It helped me really understand how it was all created.
LM: Thank you. So, just in terms of educational value: do you think an application like this could meaningfully support early literacy, listening comprehension, or imaginative engagement for young children?
GG: Oh, most definitely. Especially in areas where learners struggle with those skills, this kind of tool could make a real impact. It can improve listening skills for sure. But like I said earlier, having two versions, one with visuals and one without, would help. Because some learners are more visual, others auditory, some more analytical. So, it's good to cater to different learning styles.
LM: That makes sense, yes.
GG: Exactly.
LM: What do you think about the current structure? Where the user selects a setting, then gets relevant characters based on that setting, and finally a theme? Does that setup seem like it encourages creativity while still being age-appropriate?
GG: Yes, I'd say so. The options are solid. And at the end of the day, it depends on you as the creator, how you shape the prompts and guide the generation process. The system you've built gives enough structure to be useful, while still being flexible enough to allow creativity.
LM: Right since this is just what I created. So, for example, I included environments like "Magical Forest," "Desert Oasis," "Snowy Land," and "Crystal Cave." Then I have themed characters: fairies, elves, tree guardians, genies for desert scenes, or an octopus inventor for underwater worlds. And depending on the environment, different characters appear, like a penguin explorer or a polar bear cub for snowy settings. Others, like a magician or animal companion, can appear across multiple settings.
GG: Great. Next question?
LM: Do you think a tool like this could align well with curriculum goals or literacy frameworks? Would teachers be able to use this in classrooms without needing extensive training?
GG: Yes, it definitely has potential. Especially in contexts where students face specific challenges.
However, the effectiveness would depend on how well teachers are trained to use the tool. If they're not confident or don't see its value, they might only try it once and then set it aside.
But if the app includes different versions or difficulty levels, it could adapt to a range of ages and learning needs.
LM: That's true. I was focusing on ages five to eight, which is generally considered the early reading stage.
GG: Yes, but there's also value in having content for even younger children, pre-reading level.
That's why the government here is investing in early childhood development. There's a huge gap. Some communities don't struggle with reading and writing, but the majority of the population needs extra support. Have you read any of the research on reading with meaning? It shows that grade four and five reading levels are shockingly low.
LM: Yeah, and these days most kids scan content instead of really reading it, especially online. We were taught about that too.
GG: Exactly. Unless you're used to reading on a phone or screen, you tend to skim instead of absorbing fully.
LM: Right. I read a lot on my phone now, so I've adjusted. But it's different for each person.
GG: Things are evolving fast. That's why it's so important to understand why young learners aren't performing optimally. One big concern is screen time, parents don't want their kids on screens all day, especially if it's not educational.
LM: Yes, I've heard that a lot.
GG: If you can make screen time educational, that's a huge step forward. Cognitive development, fine motor skills, following steps, these all have to develop at certain stages. Stories like the ones you've created can help kids build focus. Right now, many of them just want instant gratification. They rush through things without processing. They skip steps and just want the end result.
LM: Exactly, immediate results.
GG: That's why something like your app is useful. It forces them to slow down, listen, and absorb the whole story. You could even prevent skipping or fast-forwarding so they're encouraged to listen to the full thing.
LM: That's a great suggestion. I'll keep that in mind. One more question: since this is AI-generated, do you see any ethical concerns? Things like bias, cultural representation, or content safety?

GG: Absolutely. That’s always a risk. Many people think AI is completely objective because it’s a machine, but it’s not. These models learn from data, and that data comes from somewhere, it has human bias baked in. Most of it originates from specific parts of the world.
LM: Mainly, I'd say the dominant training data in these models still comes from a very Western perspective, unless it’s a model developed in China, then it tends to be very Asia-focused.
GG: Yeah. And it’s also about who has the funding. We’re seeing billionaires buying everything up and directing the flow of this space. There's a real risk here. As long as we humans have the final say and control, we’re okay. But if we lose that oversight, things can get out of hand.
LM: Exactly. That’s something I’ve been thinking about too. Right now, all validation in my tool is automated, readability scores, audio clarity, that kind of thing. I’ve considered AI-based validation, but in educational settings, it might be better to include some level of human validation too. Like a teacher quickly reviewing the story and approving it before use.
GG: Yes, human intervention is still necessary. We shouldn't depend solely on more technology to check the output of other technologies.
LM: At the moment, I’m using tools like the Flesch Reading Ease score, the Flesch-Kincaid Grade Level, and some basic audio quality checks. Are there other strategies or metrics you know of that could help assess content quality, especially in terms of age-appropriateness for children?
GG: Well, these things are still new, so there aren’t tons of fixed metrics. Your original idea was for parents and families to use it, but you're also targeting teachers. That brings its own challenges. Here in our context, many teachers are older and not very tech-savvy. They're used to printed books, and now you’re asking them to trust a machine to generate content.
LM: Yeah, exactly.
GG: So, their own digital literacy becomes a key factor. What you’re asking, about evaluating stories for children, does already exist in some form. For example, our libraries and education systems classify books by age and reading level. But in schools here, we’re guided by the CAPS document (Curriculum and Assessment Policy Statement). It’s quite descriptive, teachers are told what needs to be taught at each grade and when. Unfortunately, many stick to it too rigidly and don’t explore extra materials or tools like AI unless they’re more adventurous or younger teachers.
LM: That makes sense. I also wanted to explore broader use cases, for instance, adapting this for language learners or even for special education needs.
GG: Yes, it can definitely help there. It could also be a great tool for writers or students learning creative writing. You can see what the AI generates, then analyze it or rewrite parts to improve your own storytelling. It can enhance skills in writing, literature, even narrative structure.
LM: Great. I’d like to move to some closing reflections. From your perspective, what would you say are the strongest and weakest parts of this project, especially from an educational and user experience point of view, but feel free to comment on technical aspects too.
GG: I'd break it down into a few areas: technical, educational, and your prior knowledge.
You clearly have a programming background, which gives you a huge advantage. But for a typical teacher who isn’t familiar with Python or coding, it would be much harder to create or even customize something like this. If the final product were more user-friendly, say a website or app with menus and buttons instead of code editing, it would be much more accessible to educators.
LM: Yeah, I agree.
GG: Technically speaking, this kind of setup needs decent hardware, your laptop seems like a gaming model?
LM: It is, but it’s about six years old already.
GG: Still, it handles the work well. But again, that’s another barrier, teachers often don’t have this kind of machine. If this could be packaged into a cloud-based service or run on lighter devices, more people could use it. The models you’ve used are powerful, but if they can be swapped out easily, like you said, just downloading a different model or updating a component, that's a big plus.
LM: Exactly. I made the code modular, so you can swap the story, TTS, or music generation model pretty easily. Ollama supports switching models with the same prompts. There are also new open-source TTS models coming out that are already better than the one I used here.
GG: That’s great to hear. Speed and ease of use really matter, especially for teachers or non-technical users.
LM: Is there anything else you’d like to add or suggest that we have not discussed yet?
GG: Yes. The tools you’ve used, whether three, four, or more, produced a good result. Even with some glitches, it's absolutely functional. Of course, your main focus was on audio, an audiobook-style story, but having a visual version too would be valuable. Maybe you can offer both. And then there’s the question: how will you make it inclusive for the hearing-impaired?

LM: For that, I'd include on-screen text, with synced highlighting. For deeper inclusion, like sign language, I've seen some promising tech, like gloves that automatically translate sign language into speech. But AI-generated sign language via video isn't quite usable yet. Hands and fingers are really hard to create properly.
GG: And for the blind? They could hear properly of course but there are deaf and blind individuals.
LM: Audio covers that need well. But for those who are both deaf and blind, I'd need to look into Braille support. Braille printers do exist, but they're expensive. One idea might be a reusable Braille-like sheet with mechanical pins, each letter formed by raised dots that can shift as the user advances through pages. That could keep it more accessible.
GG: Yes, and many of these ideas are still developing, but they're important to consider.
I must say, what you've built is impressive. It took time, sure, but the result, a 4 to 5 minute story, is ideal. You could also think about offering short, medium, or long versions to suit different attention spans.
LM: Yes, that's a good idea. Let users choose the story length.
GG: And even outside of storytelling, I can see this being used to explain complex topics.
You're essentially using AI as a teaching tool.
LM: Yeah, I sometimes do that too, if I don't understand something, I'll just ask ChatGPT, "Can you explain this to me?" I know ChatGPT also now includes voice options, and you can choose the voice you want.
GG: Right. Maybe this is a little off topic, but what about something like Siri? That platform's been around for a while, could it be used for something like this?
LM: With Siri specifically, I'm not sure about its capabilities directly. But I do think you could have Siri interface with another application. So, you might say, "Hey Siri, ask [app name] to explain this concept to me". Then you could customize the app: tell it your age, what you know, like "I'm good at philosophy and art but not great at math" and it could explain a math concept in a way that fits your background. Siri would transcribe your message, send it to the app, and read back the response.
GG: Makes sense.
LM: Also, just to confirm, how should I refer to you in my thesis? I believe it's Dr. Gamiet?
GG: Yes, Dr. Gassant Gamiet.
LM: Great. I'll list you as part of the Faculty of Education at UWC, specializing in science education and technology. I'll also add your title as president of the Robotics and Coding Club.
GG: Perfect.
LM: One last thing: do you have any objections if I contact you again for clarification or follow-up?
GG: You're welcome to. And I'm sure there are other colleagues who would also be very interested in what you've done. Especially given the opportunity this kind of project offers, bringing in those languages and stories that have often been left out due to lack of tech access.
LM: Yes, even if I don't include everyone I meet in the thesis, I'd still be happy to share my project, just to let people explore it. And if they use it, I'd just ask for credit where it's due.
GG: Have you tested it with parents who read to their children?
LM: Not yet. It hasn't been tested in real-life use, it was more of a prototype, developed from my research question to explore the idea. My nieces are honestly too energetic to sit through a story like this.
GG: Still, it would be good to test it on actual children, play a story to a 7- or 8-year-old and ask, "Was that a real person reading or a machine?" Or show it to kindergarten teachers or parents and see what their impressions are. There's real potential here, especially for reviving cultural stories that have been forgotten. Stories about fishermen, rural villages, local myths, these could all be brought back with this tool.
LM: Exactly.
GG: And while some people have negative feelings about AI, saying it might replace jobs, I see real value in how it can support learning. Children, students, even parents can benefit. And institutions can test whether this approach works.
LM: I definitely think this would need to be picked up by a development team to become a full application. I have the back-end working, but to create a polished app or website front-end is outside of my skillset. Still, if anyone wanted to use what I've built to go further, I'd be very open to that. Everything I've done is open-source. The research is there. Once the thesis is written, it could be released to the world, and maybe someone will take it further, as long as it's used responsibly.
GG: You've done wonderful work. It's a lot of effort to create something like this. I wish you all the best with writing the thesis.
LM: Thanks!
GG: And your mentor, what's their feedback been?
LM: My thesis mentor is also one of the lead instructors for the AI engineering track. As far as I know, they were happy with the research project. I got a 13 out of 20, which might not sound high, but considering some people

failed, I was more than happy with it. The technical side is what they focus on most. But the overall reception seemed positive. I've also already introduced the potential for collaboration with UWC's language department. If my school's AI lab is interested, we could potentially develop this further, like creating localized stories in Zulu or Xhosa.
GG: That's so important, especially for languages that have been overlooked. Since this is audio-based, it's easier to distribute. Unlike video, audio files take up less space. And many people now have access to phones where they can listen to content.
LM: Exactly.
GG: So if you were to do this project again, how would you approach it differently?
LM: That's actually something I'll cover in the "Advice" section of my thesis.
For one, I'd definitely plan around using a storyboard, it would help structure the AI-generated stories better. I also know now that background music isn't really feasible with current models, so using ambient loops or mood-based themes could be a good workaround.
GG: And what about the literature, have you looked at how children's stories are normally structured?
LM: Not specifically. My literature review focused mostly on AI story generation in general, longer stories and their challenges, like how models tend to forget what they've written across chapters.
I did implement structured storytelling techniques, like the three-act structure (beginning, middle, end), because I found that in a research paper and it made sense to use it, even for kids' stories.
GG: Did you consult with anyone from a literature background?
LM: No, not for this project. We were told to work independently, find academic sources and build from there. So, I looked for papers that matched what I was trying to create and focused mostly on the technical side.
GG: That's fair, but you should also consider how stories are used in curriculums. Often, they're chosen for cultural, political, or developmental reasons. Eurocentric stories, for example, are full of kings and queens. Other traditions use stories for revolution, or to reflect the struggles of working-class communities.
LM: Yeah, like French stories that were part of political movements, especially with revolutions.
GG: Exactly. And even in educational stories for children, there's usually a purpose, developing cognitive skills, inspiring imagination, teaching empathy. When kids read or listen to stories, they picture things. In audio form, especially, their imagination fills in the gaps. So even the setting matters, is this forest in Germany? Is the underwater story set in the Pacific or the Atlantic?
That's why context is important. Teachers should be able to guide the story generation to reflect their culture or region.
LM: Absolutely.
GG: So, does your app allow for that kind of customization?
LM: Yes. You can change the metadata and the prompt instructions, so the story can take on different tones, cultures, or locations.
GG: That's good.
LM: Okay, I think we're done with the interview then. Thank you again for your time.

Interview transcript – Rudolf Visser

Interview with Rudolf Visser, a data engineer and analyst at Data4 (internship company).

LM: Okay, everything seems to be working, so I think I'm ready to start.
RV: Sure, go ahead.
LM: Alright, let me first start with the introduction. So, I sent that info packet out, is there anything you'd like clarification on regarding the project I made, the proof of concept? Or was everything clear for you?
RV: Right, okay. It was the project about AI story generation, where you specify the parameters and ask it to generate the story, read the story, and add music to it.
LM: Yes, correct.
RV: I got all that, yes.
LM: Great. Then let me ask about initial impressions. Based on what you've seen, do you think the core concept of the AI storytelling application, with story generation, text-to-speech, and music, makes sense in an educational or industry context? Did anything in particular stand out to you?
RV: Well, you see, the only way for me to really judge it was from the little video you sent, which ran through it briefly. For me, it's about the user interface. It depends on who the target market is, is it going to be used by a teacher or someone who isn't very computer literate? Because it didn't look like a polished product. It looked more like a Python IDE-type interface. But, I mean, the concept is great. For someone having a story read aloud by a computer, that could be very useful in certain scenarios.
LM: Yeah, I know, definitely. This was just a demo proof of concept, made in about four weeks. If I were to publish this as a proper application, I'd definitely develop a proper interface, likely a mobile app, or possibly offer both a web and mobile version. For the demo, I used Gradio because it's an easy way to showcase functionality, and it has some built-in components that helped with that. We do cover basic front-end development in my major, but once you specialize in AI engineering, the focus shifts heavily to back-end, which is why the front-end here is quite basic.
RV: Okay, sure.
LM: Alright, and then from a technical perspective, what did you think of the code pipeline I created? I tried to make it modular so you can easily adjust or replace models for different components. Do you think there's room for improvement or optimization there?
RV: When you say pipeline, do you mean the way everything fits together technically?
LM: Basically, yes. You can see that more clearly on the GitHub repo, I'll just drop the link here in the chat. All the code is in the app folder, and I created separate Python files for each step of the process: story generation, TTS, music generation, and audio combining. Then in app.py, there's a pipeline that goes through all those parts, that's what I meant when I said pipeline.
RV: Oh yes, that's the one you showed in the video demo. Of course, coming from my background in ETL and data engineering, "pipeline" means something a bit different. But anyway, let me just quickly open this up. Yes, the way it's structured looks quite flexible.
LM: Yeah, that was my hope too.
RV: I think it definitely makes it easier to make changes later on, it's not too hardcoded. You can add new options, extend things. For example, I saw there was a part where you chose the environment or setting for the story. Maybe you could add even more things in the future, like specifying an intended age group, or even a language. Is that kind of thing possible?
LM: Yes, all of that is stored in a JSON file called the front-end metadata. The current setup makes it really easy to adjust. You could add more settings, characters or themes, and definitely extend it with age categories or other user-defined input.
RV: And just out of interest, can you also change the voice, like use a male voice instead of a female one?
LM: Good question! So, the TTS model I'm using is called MeloTTS, it's fully open-source. You can choose the voice, but in this case, I run it in a Docker container. You can restart the container with a different voice, but most of the options they offer are female voices. It's not the most diverse TTS in terms of voices, unfortunately. There are a few accent options though, US, UK, and Australian for example.
RV: Maybe even South African? It's actually recognized as a distinct English accent in some systems.
LM: I don't think that one was included, sadly. But if you switched to a different TTS model that supports it, you could definitely use a South African accent. I just stuck with the default UK English voice for the demo. Initially I had planned to support some multilingual elements, kind of like what Dora the Explorer does, with simple vocabulary words swapped out in another language. But I had to scrap that idea. It was too tricky to find a story generation model that could reliably support that, and even when I thought of hardcoding vocab swaps in Python, the TTS side became a problem, it was hard to find a solid, open-source multilingual TTS model. So, in

the end I kept it all in English.
RV: Sure. If I can just ask about prompting, as you know, with generative AI it's usually about entering a prompt. Did you regenerate that prompt based on the user's inputs, like character selection and the other things? So, you're not allowing them to just type in a full sentence as a prompt?
LM: So, the way it works is that I use very specific prompt engineering for the text generation model I'm using, which is Llama 3.1. The prompt includes metadata such as character names and descriptions. The story itself is generated in three parts, I have separate prompts for the beginning, middle, and end of the story. I can quickly show one as an example, this is the prompt for the beginning of the story. I specify that it's a storyteller writing the beginning of a children's story, that it should use simple, clear language suitable for the age group, and I provide the chosen setting, characters, theme and their descriptions. Because it's the beginning, I want it to introduce the world and the characters, have them interact a bit, and present a problem or issue they'll need to solve, ideally with a little mystery or surprise to lead into the middle part.
RV: Ah, okay, I see. So, you're obviously making it a lot more specific than just a general prompt like, "Tell me a story about a cat that walks in the forest," or something like that. It's more about setting up the whole plot, how the story should go, with a beginning, middle, and end. Was that one of the requirements of the project or was that your own idea to do it that way?
LM: That was my own idea, but it was also based on some research I did. I found a paper that talked about using the three-act structure and how that helps improve coherence and overall story quality when generating stories with AI. The starting point for the whole project was my main research question and some sub-questions I defined. The main question was: *"How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?"*. But the background music turned out to be less feasible, both in terms of time and resources, and also because of limitations in the current models. For example, the model I used, Facebook's MusicGen-Small, can only generate three-second clips. So, I ended up using those more as little transition pieces during the narration, instead of having a continuous background track.
RV: Yeah, I get that. Of course, you could always mix a voice track and a music track afterwards if you wanted, but then it wouldn't be automatic. That would require some human intervention. Or is that something AI could do too? I mean the actual mixing of tracks?
LM: There might be options for it. I did come across some interesting music models with live demos that seemed to have potential for that. But in my case, I didn't just want a constant background track. If something in the story changes, like if there's a moment of suspense or a surprise, I wanted the music to reflect that change. Which, of course, made things way more complex. Because then I'd need timestamps in the narration to know where the tone shifts, so I'd know when to change the music. It just became a bit too much for this project, especially with the limited time, I had about four weeks total, and at least two of those were spent just refining the story generation. I went through a lot of different prompt versions before I found something that worked fairly well, I'd say about 90% of the time it now produces decent results. Occasionally, though, you'll get odd things like repeating a character's description or a story section that ends with "to be continued" even though it's just the beginning. But overall, it's quite consistent now.
RV: Interesting. So, is it different every time? Like, if you run the same thing twice, the same exact metadata, do you get different stories each time, or does it always generate the same one?
LM: If you run the exact same metadata, same setting, characters, and theme, you'll get a different story each time.
RV: Ah, okay. That's interesting.
LM: It might reuse some elements, for example, the model seems to really like using the name "Luna" for female characters, but I just provide the metadata and tell it to create a story with that. The rest is up to the model. And then each part of the story, beginning, middle, and end, gets passed into the next prompt. So, the middle prompt also includes the full beginning output as input, and I tell it to continue from there.
RV: Yeah, okay. Makes sense.
LM: So, next I was wondering, especially since you work more with cloud services, if you were to imagine deploying an application like this to the cloud at scale, what kind of considerations come to mind? Do you think the current architecture would be adaptable, or would it need major changes?
RV: Hmm, well I'm not entirely sure, since I'm not a developer myself. So, I don't know exactly what components would be required to publish something like this on the internet. But from what I do know, things like this are usually scalable. It would really depend on the expected audience. Say you wanted to roll it out to all the schools in Belgium, then you might be looking at 10,000 users at the same time. Typically, when you allocate resources on Azure, you have to specify the expected number of concurrent users. Based on that, it adjusts your compute and storage allocation. You can set it to be scalable, but of course that tends to be more expensive. That said, I definitely think the cloud environment is well-suited for something like this, even just running it from a browser.

And these days development environments are super flexible, so you can create something that also runs on a mobile phone. That’s just a different front-end with a smaller screen, but it talks to the same back-end, gets the story and audio and plays it back, whether that’s on a phone speaker or a computer. So, I’d say the cloud is the way to go. Of course, you'd also have to consider things like intellectual property, or if there’s any copyright-sensitive material involved. I don’t know if you looked into that for the models you used, is that something to worry about?
LM: I did make sure to include the licenses in my code. MeloTTS is fully open-source, you can even modify and publish it. The music model has an MIT license, and the Llama model has Meta’s specific license. I’ve included all those licenses as txt files, and they’re downloadable from the front-end. I also list which models are being used at the top in Gradio, and at the bottom of the app there is the download link. I wanted to make sure everything was properly covered in terms of copyright and licensing.
RV: Yeah, that’s important. Especially if you ever want to go public with it, make it available to others, or even turn it into a commercial product. Then the cloud providers will usually charge based on user count or usage, so that’s another thing to keep in mind. But still, it’s so much more flexible if you run it on a cloud platform, I think that’s definitely the right direction.
LM: Alright. Let me move on to the next question. These are more focused on the educational aspects now. One of the main goals I had in mind for this application was that it could be used by young children, probably with a very different UI, of course. I’d want something that uses simple icons they can tap on a tablet, maybe with integrated text-to-speech explaining what each part does. Like, when they hover or tap, it says, “This is the story section” and maybe flashes or changes color, then lets them choose characters, etc. But it could also be used by parents or preschool teachers. The idea was to promote reading engagement and support listening comprehension for young kids.
RV: Well, I wouldn’t necessarily call it reading, it depends. Are you going to show the text for them to read as well, or just the audio for them to listen to? Otherwise, it’s more of a listening experience than a reading one.
LM: My idea for a full application would be technically both. It would have narration, but I imagine something similar to karaoke, where the text is highlighted or followed along visually. So, the child would also see the story as it’s being read.
RV: Ah, I see. So, both text and voice, that’s great for teaching children. Yeah, that makes sense to me. I agree it would be most beneficial, especially for children when there isn’t a parent or teacher available at the time to help them read or learn.
LM: Exactly, that’s also what I remember from when I was learning to read. It was often a “read together” experience, with a parent or teacher reading aloud, and I’d follow the text with my finger on the book. This aims to be a digital version of that same experience.
RV: That’s good. Makes sense.
LM: Okay, next, do you think an AI-powered storytelling application like this would fit well in the educational tech or content creation market? Do you think this is something that would be marketable?
RV: I would think so. I’m not directly involved in the education space, so I don’t know what all is out there already, but not knowing the competition, I’d still say yes, it definitely has potential. These days, many parents just don’t have time to regularly read to their children. And in schools, especially in South Africa, class sizes are often very large. There’s little opportunity for one-on-one attention, particularly for children with learning difficulties. With an app like this and the right device, those children could carry on independently and spend more time developing their reading skills.
LM: Alright. The next question is a bit more about the business approach. This proof of concept is currently fully open-source, and if I imagine it as a full application, I’d still like to keep it open-source, mainly because it’s aimed at education and child development, which I think are important. So, I wouldn’t really want to put a price tag on it. Do you think an open-source version could be appealing? Or would a hybrid or licensed model be more practical if this were to go into the real world?
RV: From what I know of shareware and open-source models, what usually works is offering a light or limited version for free. That gives people just enough to try it out, start liking it, and see the value. Then if they want extra features or full access, they’d pay a license fee. I’d suggest a hybrid model like that. Another common approach is offering the full version for free, but only for a limited time, like a trial for one or two weeks. After that, users either pay or switch to a basic version with limited functionality. I think that kind of model could work well from my own experience. It gives people a chance to try it out first, and if they enjoy it, they’re more likely to pay to unlock everything.
LM: Okay, next question, still on the business side. Do you think investing in generative AI is a good idea, especially compared to simpler template-based tools? Would you say that the AI element is a strong differentiator?

RV: Yes, I definitely think so, especially because of the machine learning aspect of AI, where the system can learn and adapt over time. If it were just a predefined template or hardcoded system, then you'd always get the same story every time. But with AI, it's different each time, more engaging and interesting for the user. I also think the range of possibilities is much broader with AI compared to a rigid, non-learning model.
LM: Yeah, absolutely. If I do imagine this as a full application, I would definitely include some kind of feedback mechanism. Something simple for children, like a five-star rating to show how much they liked the story. But for parents or teachers, there could be a more detailed feedback form, maybe something like "I liked this story because…" or "I didn't like this part…". That kind of feedback could then be used to improve and refine the models over time.
RV: Definitely. You could even ask the child things like, "Who was your favorite character?" or "Would you like to see more characters like this one?". If you already have a character list, you could collect statistics, such as out of 100 children, 80% chose this character and only a few chose others. That kind of usage data could then help the model learn what's more popular or appealing.
LM: Yes, exactly.
LM: Alright, next, do you think there might be any ethical concerns with AI-generated children's content? Things like bias, cultural representation, or content safety?
RV: That depends a lot on which model you're using, and what data it was trained on. But based on what I've seen from your demo, where you have a lot of control through prompt engineering, I think the risk is a lot lower. You're not just giving it freehand prompts, you're guiding the generation quite specifically. So, for example, you can define the race or context of the characters instead of leaving it up to the model to assume or guess. That level of control can help avoid some of the common ethical issues. With the kind of setup you're using now, I don't think it's a major concern, compared to more open-ended generation systems.
LM: I can also share that I do have a check for prohibited words. So, if certain words appear in the story, it is automatically told to regenerate it. These include things like hate, prejudice, violence, weapons, adult themes, self-harm, or mental health issues, and crime as well. Especially since this is meant for young children, I try to be as careful as I can to make it safe with the current setup.
RV: And you also make it a good story, like none of the characters die and it always has a good ending?
LM: Oh definitely. That's one of the points in the ending prompt as well. I tell the AI to make sure the story has a happy ending.
RV: Okay, yeah, a happy ending.
LM: Do you think it might also be good practice for safety and quality to include some human validation as well, and not just rely on the automated checks? Like if this is used in a classroom or by a parent, that the adult first checks the story before it's played to the child?
RV: Well, no, I don't think so. I think those checks and balances should be built into the system itself. Like you just showed, the list of restricted themes and words, I think that already covers the moderation needed. Sometimes, you just want to let the child explore and do their own thing without an adult needing to check everything. So, if you make sure the system works within the boundaries of what a school, for example, would prescribe, then I think that's enough. Of course, the school should understand what the system does and what parameters it follows. That way, they can decide if it fits their needs. Some people might only know about AI from what they hear on social media, where there are horror stories of AI going wrong. So, it's important to reassure teachers, this app only creates nice stories, nobody dies, nobody swears, nobody does bad things. That would likely reassure them.
LM: Yes, I understand. I actually read a paper that described a story generation setup using two models, kind of like a GAN system. One model creates the story, and another model, which they called the "action discriminator" checks in regularly with the generation model to see if anything needs to be changed. So, in a full application, there could definitely be a version where a second AI model handles validation, not just hardcoded keyword checking, but true AI-based validation.
RV: Oh, I didn't even know that was possible, that one AI could check another AI. That's very interesting.
LM: Do you think something like this AI storytelling could include more emerging technologies too? Like real-time interactivity, more personalized narration, maybe even stories where the child is the main character or adding AR or VR components for a visual experience?
RV: Actually, that makes me think, there's already a big industry around audiobooks, right? Mostly it's humans reading and recording them. So, this could actually compete with that space in a way. For example, I was just thinking about elderly people who live alone and don't have many visitors. They could use this to have a story read to them, or even better, generate a story they'd like to hear. Something based on memories or themes from their youth. So, in that way, this could be an alternative to traditional audiobooks.
LM: Yes, definitely. If I imagine it for an older audience, you could even have a very simple speech-to-text

function, they could just say what kind of story they want, and the AI could use that input along with some preset templates to create something completely new.

RV: Yeah. Actually, that reminds me, my wife reads on a Kindle, and many websites use similar systems. They learn your preferences over time. So, if you like a certain author or type of story, it recommends something similar next. So maybe you could also ask the AI, "Tell me a story similar to Lord of the Rings" and it would generate something in that style. Of course, copyright could come into play there, but conceptually it's possible. Users could choose a genre or a type of story, and I'm sure AI could come up with something close to that. And it's like what we already see in the music industry, when you listen to a few tracks, the system quickly learns what you like and gives you something similar next. It's already widely used, so I'm sure the same principle could work for storytelling.

LM: Yes, absolutely. Then also, thinking ahead, do you think future text-to-music models that allow for longer, more coherent music generation would significantly improve the overall storytelling experience?

RV: Well, it depends. Not everyone likes background music while listening to a story. Some people just want the words, just the story itself, no music. So that's probably something where the user should be given a choice.

LM: Yes, definitely. I imagine a simple toggle option, such as a "Yes/No" button labeled "Would you like background music with this story?" That way, users can choose for themselves.

RV: Yeah, exactly.

LM: And do you think there might also be a demand or educational value in generating and narrating stories in multiple languages or dialects? For example, there's already been some interest here from the language department at UWC, they were curious about using this to create stories in native African languages like Zulu and Xhosa.

RV: Yes, definitely. That's a really good idea. But there is one potential issue, for AI models like this, you need a device that can run it, and you also need connectivity to a server in the background. That might be a limitation in certain areas. In a third-world context, like in South Africa or other African countries, many native language speakers might not have access to that kind of technology. So, you'd need to make sure it runs on a simple or low-cost device, maybe even older phones. That way, access to the technology matches the economic background of the user. In principle, yes, more language availability is a great idea, because there are often very limited resources in those languages, especially for children from underprivileged communities.

LM: Yeah, no, indeed. If I look at something like that, I don't think people in those situations would have the resources to run this locally like I currently do or even have stable internet connectivity for it to work properly. So, in that case, it could be something like an MP3 player that already has pre-generated stories loaded on it, stories that someone helped generate and checked using the AI beforehand. Even older tech like cassette tapes could still be useful if that's what they have lying around.

RV: Yeah, that's possible. But of course, then you lose the interactivity, like being able to choose your characters or the setting and so on.

LM: True. Alright, then moving on to the closing reflections, just from an overall perspective, what would you consider to be the strongest and weakest aspects of the project so far? From a technology or user experience point of view.

RV: Since it's still a proof of concept and a very early version of it, I can't really comment too much on usability. You did provide an installation manual for how to set everything up, but I haven't tried it myself. It's hard to judge since it's still far from the end-user context. But from the back-end side, I think you made very good use of the tools available to you. If that complexity is hidden away from the end user in the final product, then that's fine. The architecture seems solid. The front-end you didn't spend too much time on, which is of course the part the user will care about the most. So, I can't judge usability yet, but I do see a lot of potential here.

LM: OK. Is there anything else you'd like to add or something you feel we haven't touched on yet?

RV: No, I think we covered everything pretty well. Nothing else comes to mind at this point.

LM: Great. Then also, for my thesis I need to reference everyone I've interviewed. What's your full official title with Data4?

RV: My official title is Data Engineer and Analyst.

LM: Got it. Do you also have a doctorate or something I should include?

RV: I've got a master's degree in physics. Is it the same in Europe as well?

LM: OK. Yeah, in Europe we also have bachelor's and master's degrees, so that fits perfectly.

RV: Yeah, sounds familiar. In my case, it was a three-year bachelor, then an honors year, and then the master's which I completed in about a year and a half.

LM: Lastly, would you be open to a follow-up conversation in case I have more questions later on in the thesis process?

RV: Sure, I'm happy to help. Even if you'd like someone to read over your thesis, like for language use and

phrasing, I can assist with that too.
LM: That would be really appreciated, it would be great to have a human perspective on the phrasing and tone. I think I'll only need that kind of review closer to the final stages, like after the feedback from school for my first deadline on the 22nd of April. The final version is due at the beginning of June.
RV: Just take note, I'll be on holiday during the third week of May, from the 21st to the 25th. So, I won't be available during that time.
LM: No worries at all. I'll make a note of that and reach out after I have feedback and make the adjustments. It would be great to have some real people check over it to make sure everything is well-formulated.
RV: Sure.
LM: Thanks again, that really means a lot.
RV: You're welcome. And I must say, you speak really good English. Where did you learn it so well?
LM: Honestly, I think it's a mix of things. Back home in Belgium, most TV shows for adults are in English with subtitles, not dubbed, which really helped me pick up vocabulary and sentence structure. Even Dora the Explorer teaches English in the Dutch version! And I remember as a kid playing Pokémon games, they don't come in Dutch, so I had no clue what anyone was saying, but I just kept pressing buttons and figured it out. I also read a lot now, and mostly in English, so I think all of that helped.
RV: Well, that definitely paid off. Thanks Lara, good luck with the rest of the thesis. We'll speak again.
LM: Thank you! I'll be in touch if anything else comes up.

Interview transcript – André Daniels

Interview with André Daniels, digital media coordinator at the Centre for Innovative Education & Communication Technologies (CIECT) at UWC.

AD: So, are you a programmer? What is your major?
LM: I am a programmer, yes. The major itself, back in Belgium, is at the Howest University of Applied Sciences. It's called MCT, or Multimedia and Creative Technologies. And then about halfway through the three years, you choose one of four branches. I'm in the AI Engineer branch, so I've learned how to program AI, how to work with it, that kind of stuff.
AD: Okay, that is like, far removed from my field. But in fact, anything that I'm doing now is far removed from what I studied anyway.
LM: Yeah, I mean, the reason I was interested in interviewing you is that even though this is a technical project, I built a little AI storyteller proof of concept application. I really wanted to get a perspective that isn't technical, something more from the business side, the application side, and even the ethical side. Did you get the info packet?
AD: Right, yes, I did. I mean, I didn't manage to go through all of it. I looked at it last night and just before our meeting. I checked out some of the stories you developed. I didn't try to install anything, my computer is way too slow.
LM: Totally fine, that's why I included the demo video. I only added the installation guide because I had to make that anyway, mostly for the more technical interviewees. Even on my laptop, which is a bit older but still a gaming laptop, it takes a while to run and uses quite a bit of disk space too.
AD: Okay, that's what I was wondering. So, tell me a little bit, just before you ask me your questions, how did you arrive at this point where you got interested in using technology for storytelling?
LM: So, last semester, back from September to January for us, we had a module called "Research Project", which is the prerequisite to the bachelor's thesis. You get your main research question, come up with some sub-questions, and then make a demo of whatever you're trying to explore. I actually got some help from my teachers in formulating the question because I wasn't entirely sure what direction I wanted to go. What I did know was that I wanted to do something with AI that we hadn't really covered much in class or in previous projects. That's how I ended up doing story generation, text-to-speech, and some music generation. I eventually landed on the question: "How can AI be used to generate and narrate a children's story for ages 5–8 with accompanying background music?" since I myself like reading a lot and I was thinking back to when I was learning to read as a child as well. I created the AI storyteller application in about four weeks, in January. Now the thesis is basically looking back on my work and also looking forward, what else could be done with this? That's also why I'm interviewing people externally, to get input from different fields and perspectives.
AD: Okay. Because this really is a completely different angle to storytelling than what I'm used to. I don't know if you've ever worked with children that age before in storytelling?
LM: No, I haven't.
AD: Okay. I've done maybe one or two workshops with kids around that age, maybe a bit older, say from 9 to 12 years old. It was with a museum that was collecting stories and wanted the kids to tell their own. A colleague from the History department who works with young people wanted to explore that. We used something like the story circle method. So, I'm more familiar with that, especially through Joe Lambert and the Berkeley Storytelling Center in California. I don't know if you're familiar with their model?
LM: I'm not, no.
AD: So, I started working with that model around 2005. It was mostly used in the context of HIV and AIDS, people sharing personal stories about how they lived with the disease. That was my entry point into digital storytelling. At the time, it was more from a technical perspective, so I was just assisting. But for me, it was also interesting because it gave me some purpose, I had technical knowledge but wanted to apply it in a meaningful way.
LM: Oh, sorry, I think you just froze for a second there.
AD: It's okay.
LM: My internet just randomly cut out. I'm on my phone hotspot now.
AD: No worries. I don't know if there's anything specific you wanted to ask?
LM: I do have a list of questions. If there's anything you don't feel you can answer, or if it's not in your area of expertise, that's totally fine. Let me just quickly check something. Okay, I think my internet is back, let me switch off the hotspot. All right, everything seems fine now. No idea what caused that, but transcription is still running fine.
AD: Good.

LM: Let me start with this: based on what you've seen of the project, is there anything you'd like clarification on? Or do you feel like you have a good grasp of the idea and the concept?
AD: Yeah, I mean, other than the scripting and obviously the programming jargon that I'm not familiar with, I think I get the premise. It's a simple interface with different themes or filters, and then the AI generates stories based on those. And it makes use of a large language model. What's also interesting is that you're using different platforms, like something from Facebook for the music?
LM: Yes, exactly. So, a very quick overview: the front-end is basic, and you select a location or setting for the story. Based on that, a filtered set of characters becomes available. I did this to avoid illogical combinations, like no mermaids in the desert, for example. Then you choose a theme, like friendship or teamwork. All of that information is sent to the large language model, in this case, Llama 3.1 from Meta. The story generation itself is split into three parts, and each part has a specific prompt. For the beginning, the prompt instructs the model to introduce the characters, set the scene, and begin the conflict or adventure. The middle builds on that, develops the conflict, and so on.
AD: So, all these prompts are happening behind the scenes?
LM: Yes, those are all behind the scenes. And of course, with the ending prompt, I also make sure to include something like "give it a happy ending," because they're children's stories, you always want a happy ending. Usually there's also some sort of moral lesson or a little knowledge tidbit. I remember that myself from when I was younger. Fairy tales often include a simple message or moral. All of that gets passed to the text-to-speech model, which narrates the story using a simple, open-source voice. Originally, I planned on having full background music, but the current text-to-music models only generate very short clips. The one I'm using right now can only produce three seconds of music at a time.
AD: Yeah.
LM: And because of time and resource constraints, especially since everything runs locally, it shifted from background music to more like little transition music clips. So, at the beginning of the story there's a short piece, another one when the middle starts, and again when it ends.
AD: Yeah. I noticed that. I was thinking, okay, technically speaking, this isn't background music. But as you clarified, it's more for transitions between scenes.
LM: Exactly. And on the front-end, once everything is done, it takes about 8 to 10 minutes on my older local setup with a 2060 Nvidia GPU, you get the full story displayed in text with an audio playback option for the narration.
AD: Alright. And the voice you're using, I think it's a female voice, is there variation? Like can you choose male or female voices?
LM: From that specific model, called MeloTTS, it was mostly female voices. The only real variation was in accents, like American English, British English, Australian, that sort of thing. I just experimented with a few on Hugging Face's demo page and picked the one that consistently pronounced things correctly. I also thought it was pleasant to listen to, and I think that matters, especially for children. The speaking speed is also a bit below the standard, 0.9 I think, just to slow it down and make it more listenable.
AD: So, to answer your earlier question, yes, I think I understand the basic principle behind it. Actually, it reminded me a little of something, I don't know if you've ever worked with Photo Story?
LM: No, but I have heard of it. I think that one also includes visual elements, right?
AD: Yeah, it's more visual. It needs more input, but it's basically like a wizard that takes you through different steps. You just add in your features, and it automates transitions. I haven't worked a lot with digital stories that are just audio, essentially podcasts. Most of my work has visuals, but I've seen people doing similar things with only audio, like you are now.
LM: Yeah.
AD: So, it reminded me a bit of that. I'd almost describe it as an automation, but with AI in the background doing all this sampling and piecing things together.
LM: Yeah, exactly. If I imagine this as a full application, my idea was to have something like karaoke screens, where the narration happens, but the text also gets highlighted as it's spoken. I remember learning English that way. In Belgium, most non-children's content isn't dubbed; it's subtitled. I remember watching shows, hearing the English, and reading the Dutch subtitles, realizing, "Oh, umbrella means paraplu." That really helped me.
AD: Okay.
LM: So that's part of the idea behind it. But if this were a real application, I think visuals would be really useful, especially for younger kids. More attention-grabbing.
AD: Yeah. So, I was going to ask, especially regarding the age range, how does the AI know what's appropriate for a five-year-old versus a nine-year-old? Like, could there be an option where a parent selects the child's age and the story adjusts accordingly?

LM: That would definitely be a great addition, but currently, it does not have that. This is what my beginning prompt looks like, for example. I specifically tell the AI to use simple, clear language suitable for children aged five to eight. I looked into early reading development and found that range to be a common reference. I also run a check after the story is generated for prohibited words, like “knife” or anything too adult. Here’s the list I use. It includes terms related to violence, adult content, or negative traits, like “loser,” “idiot,” “dumb,” or “stupid.” Stuff you don’t want a child to hear or learn from a story. There’s also a Python library that analyzes readability using something called the Flesch Reading Ease scale, which rates from 0 to 100, where lower is harder and higher is easier. It also calculates a Flesch-Kincaid grade level. I try to keep it equal or under grade six. The stories may be slightly above the reading level for a 5–8-year-old, but since it’s narrated and not something they read on their own, I think that’s acceptable. All right, next question then. Based on what you’ve seen and all the info, do you think that the core concept of something like AI storytelling makes sense in an educational or industry context? Does anything about it really stand out to you?
AD: Yeah, I was going to ask you about that, because like I said, my first thought was, okay, AI coming up with an idea. And I was just wondering, in terms of how I work, I usually approach storytelling from the perspective of people telling their own stories. So, the idea comes from the person. I was going to ask you, thinking back to your own experience with storytelling, what was the most exciting part? Was it the story itself, or the fact that someone sat with you and told you the story?
LM: I think it was the fact that someone was there with me. Once again, If I imagine this as a full application, one thing that AI can now do quite nicely with text-to-speech is voice cloning. So, I’m thinking about how, in today’s world, both parents often work, and there isn’t always that much time for the child. In that case, it might be really nice to have the option to record 30 seconds of mom or dad speaking and then use that to generate the entire story in their voice.
AD: Yeah, that reminds me of something I did with my daughter using just a cell phone. We’d read the story and record it, and at night she’d play it back. But in most cases, of course, we were there reading it. The recording was more for our convenience than for her. I think from an educational perspective; it depends on what you're planning to use it for. At the end of the day, it’s another tool. The real question is how you plan to use it. The context matters. Stories aren’t necessarily meant to stand alone. They can be teaching tools, sure, but often they’re designed to spark conversations or further discussions. So, this could be a springboard into reading, learning to read or developing a love for reading. I often say that the stories we develop are not stand-alone content, they are designed to enhance interaction. An important thing to ask is whether this becomes a replacement for an adult reading to a child, or whether it works alongside a person teaching the child to read.
LM: Yes, I definitely see this as a tool to be used in tandem with others. One aspect I really like about what I built is the customization. You can choose what the story will be about. If it’s developed further, you could even have the child featured in the story. If a child is really into dinosaurs or has a favorite pet, that could be integrated into the story. It’s about making the story personal to the child, which could help with interest, especially in early reading or even listening comprehension.
AD: Okay.
LM: If this were used in education, the teacher would still be there to assist. But let’s say some kids in the class are slower readers and the teacher can’t focus on them all the time. They could use this tool on their own with some support. Or, like I mentioned, if a parent doesn’t have time that night to read to their child, they could still provide a bedtime story using their own voice cloned into the narration. I don’t see this as AI replacing people, more like a tool to help them.
AD: Yeah, I agree. I’d also see it more as a companion than a replacement. And that idea of customization is great. A child at five or six has their own imagination. How do you bring that into the story, either by creating something new or tweaking an existing story? Like you said, what if the child doesn’t like how the AI ended it? Maybe you let them bring in their own pet by just pressing a button, entering the name, and regenerating the story with that change. That sort of interactive customization would be really nice.
LM: Like I said, this was very much a proof of concept, but in a full application you'd have a library of generated stories. You could give feedback to the AI like, “I liked this part,” or “I didn’t like that.” And you could probably even say, “Give me another story with this character, I really liked them,” you know, that kind of thing.
AD: I was just telling my daughter yesterday, she’s also really into movies, that nowadays, every day on TikTok there’s someone showing off a new feature you can run on your desktop. The biggest shift for Hollywood might just be that all you need is a good story. Because soon, people will be able to create entire productions right from their desktops. If you think about it, the real differentiator isn’t the tech anymore, it’s the idea. And Hollywood stories lately aren’t always landing. Like that new Snow White movie, it didn’t perform well.
LM: Yeah, the live-action Snow White, I saw videos about that.
AD: Right. So now, if someone has a good story, they can just use an AI application to bring it to life. That’s

where we’re going, AI handles the production.

LM: I definitely agree. Though I will say, AI isn’t great yet at writing long stories. The ones I generate are between 800 and 1,300 words. If it gets too long, the model starts forgetting earlier parts of the story and things become inconsistent or illogical. A full movie script, with long context and complex narrative arcs, that’s still a challenge. But for short, contained children’s stories, it works really well, especially with refined prompts.

AD: Okay.

LM: So then, do you think an application like this, once it's fully built, could meaningfully support early literacy, listening comprehension, or even imaginative engagement for young children? Do you think it could have a positive impact?

AD: Absolutely. I think it’s especially powerful because it’s still text based. So, the listeners still get to conjure up their own images in their minds. Normally, for example, when reading a bedtime story, you’d have pop-up books or visual cues to capture the child’s attention. But even with audio only, it becomes a springboard. In fact, you could use it as an assessment tool. Imagine giving kids a series of generated stories and then asking them to select pictures that match the story, or even draw their own. You could build from there, layering their interpretation on top of what the AI created. It’s definitely a strong engagement point for development.

LM: Also, how effective do you find the current selection mechanism? So first, you choose a setting, then based on that, you choose up to three characters, and lastly a theme. Do you think that encourages enough creativity without being too restrictive?

AD: That’s a tough one to judge. If you’re thinking of an adult creating the story for the child, then it’s quite basic. But if children are the ones using the app directly, it could feel rich to them, especially if it's built visually with icons rather than just text.

LM: Yes, the final version would likely be a tablet app with big icons they can tap on to make selections.

AD: Right, that makes more sense. In that case, I think it could definitely support creativity. You’re giving them the elements of a story and helping them learn how stories are structured. It could even teach them how to build a coherent narrative themselves. So yes, I think it’s got potential to help them develop storytelling skills, not just consume stories.

LM: Are there any features or changes you’d suggest to the current demo concept to make it more effective or more engaging? One idea I had for a full application is that, since most children that age can’t read well yet, there could be voice guidance. Something like: “Welcome to the storyteller app. First, choose your setting.” Then it would flash or animate that part of the UI. “Now, pick your characters,” and so on. It would walk them through the process.

AD: Oh, you mean like a voice prompt to guide them through the process?

LM: Yes. If a child is using this themselves, they could still navigate it with voice guidance. They wouldn't need to read any text. Just a voice and some clear icons helping them.

AD: Off the top of my head, I’m thinking about some mobile apps I’ve used, especially those for language learning. It might be useful to have an initial setting where the user can select their age or experience level, something like beginner, intermediate, or advanced. That way, the experience could be tailored accordingly. And like you said, there’s already the option of choosing an accent for the TTS. That could be expanded to different languages, and even subtitles or captions could be added. Subtitles are definitely useful, especially for promoting multilingualism or supporting reading development.

LM: Yes, absolutely. In the beginning, I was thinking about multilingual support, kind of like *Dora the Explorer* style, having basic vocabulary in a second language. But I had to scrap that idea because multilingual text generation and especially free, open-source TTS for other languages are really limited right now. But it would be great at that young age when kids can still easily absorb another language, just having some vocab or even the whole story in a different language.

AD: Yeah.

LM: I actually had a conversation about this with Gassant Gamiet and the African Languages Department at UWC.

AD: Oh, okay.

LM: They were interested in the idea too, especially around adapting it for languages like Xhosa or Zulu. I got them in touch with my university, Howest, because we also have an AI research lab. So hopefully something can come from that. It would be amazing to see underrepresented languages better supported in AI-generated content, especially for storytelling where local language presence is minimal.

AD: Absolutely. I’ve been using digital storytelling to encourage multilingualism. I build in subtitling and audio dubbing to extend the usefulness of the stories. I actually just met with someone yesterday who reminded me, sign language is also a national language in South Africa now. I’ve done some stories with sign language interpreters as well.

LM: Oh, yeah. That’s really cool. I remember seeing a project some years ago where someone developed gloves

that interpret ASL movements and vocalize them. It wasn't AI-based, more sensor-driven and mapped to a predefined vocabulary library, but still fascinating.
AD: That sounds amazing.
LM: Yeah, definitely worth looking up. And here in South Africa too, I've heard from someone who taught their daughter sign language when they were little, just basic signs like "hungry" or "tired", and it helped with early communication before they could speak.
AD: Yeah. Actually, a few years back, before AI was so widely available, some students from the Computer Science department at UWC worked on an application for South African Sign Language. They were using camera input and training a system to recognize gestures.
LM: That's awesome. Just quickly, I actually found the clip about those ASL gloves, it's from 2016, quite old, but really worth a watch. I'll send it to you after the interview.
AD: Oh, great.
LM: Back to the topic, do you think a little storytelling app like this would align well with curriculum goals or literacy frameworks? Do you think it would be practical for teachers to use it in classrooms if they had a simple handbook or short training?
AD: Yes, I think your info pack already shows that it can be done. From what I saw, the interface is simple enough for a non-technical user to understand and use. It could absolutely work in a classroom. But more importantly, it's how it's used that matters. You want learners to be actively engaging with the tool, not just passively listening. So, while teachers could use it as an instructional tool, the real value would come when learners start using it to develop their own stories. A simple train-the-trainer approach would work well here, once a teacher learns the tool, they could guide learners step by step. It reminds me of Photo Story, a very linear process where you didn't need deep technical knowledge, just the ability to follow the steps and prompts.
LM: Yes. And do you also think there could be ethical concerns with AI-generated children's content? Things like bias in training data, cultural representation, or safety of content?
AD Yeah, I think you've touched on a lot of important points already. What's acceptable varies so much depending on culture. Would the AI even know what cultural setting it's generating in?
LM: It doesn't currently, I can tell you that much.
AD: Yeah, that's obviously a key concern. And I suppose that's why the responsibility will always lie with the end user, to determine what's appropriate. There should probably be an option to regenerate a story if it contains something off. For example, imagine the AI generates a story in an Indian context, and it includes something like eating or slaughtering a cow. That would be highly inappropriate given the cultural context. So, it's not easy to anticipate every issue in advance. Like with ChatGPT and other AI tools, users will need to be able to review and judge whether something is ethical or not, especially since those norms shift so much between cultures. And because your application currently doesn't include images, that does simplify things slightly. You're minimizing the risk of problematic visual representation. Still, you could bring in human engagement through something like printed boards or picture cards. For example, while the narration plays, children could select images that best represent the story they're hearing. That invites interaction and makes the experience more immersive, without needing full-on visual integration.
LM: Definitely. I remember one story where the AI described a mermaid's hair as "a golden river," which could suggest blonde hair. But many characters are never described physically at all. So, depending on where the child lives, they might imagine the character as someone who looks like them, Caucasian, Black, Asian. Kids naturally gravitate toward seeing themselves represented in the stories they hear.
AD: Yeah, absolutely.
LM: I do have some basic automated checks in place, but in a real-world context, I'd imagine this being used with a parent, teacher, or other trusted adult present. Maybe even with some kind of parental controls. Like, the adult could read through the story first and give it the go-ahead. Approve it for narration. That hybrid approach, blending human oversight with AI tools, feels like a healthy balance.
AD: Yes. I was actually going to ask, since you mentioned the pacing and inflection of the narration, would it be possible to edit the text before the narration is generated? So, the story text shows up first, then once it's approved, the narration starts?
LM: That could definitely be implemented. I could see a step in the process where, after the story is generated, the adult user reviews the text. Once they're happy with it, they could approve it, maybe even behind a simple lock, like a basic math question only an adult would be able to answer. Once approved, the TTS would generate the audio. And in a full application, the text-to-speech model would definitely be upgraded to something more expressive. Some of the better TTS models today are incredibly good at sounding natural.
AD: Yeah. We recently worked on some screen recordings using AI-generated voices, and it was really tough to find one with a South African accent. We did find one eventually, called Love Voice. The biggest challenges were

always with inflection and pacing. The voice would stay flat the entire time, and all the pauses were the same length, which just doesn't work well for storytelling. With longer stories, it becomes noticeable and a bit monotonous.

LM: Yeah, definitely. I've tested a few browser demos myself, and some of them do allow you to select emotional tones, happy, sad, neutral, etc. I could imagine in a full-scale app you'd have one voice for the narrator, and then different voices for each character. You could even add emotion tags in the text, like putting something in brackets that says [angry], so the voice knows how to say the line. Some advanced models could even analyze the text and add smart pauses or tone shifts automatically. That would make it really dynamic and engaging.

AD: Yes, for sure.

LM: Also, do you think if future music models allowed for longer, coherent background music, would that significantly enhance the storyteller experience? Or do you think it would just be nice to have rather than essential?

AD: Yeah, I mean, when it comes to stories, adding music is very subjective. It depends on how you're telling the story and what sort of tone you want to set. Normally, we start with a podcast version, and I always say that if a story works well in audio alone, if it can move or engage the listener, then it already stands strong on its own. Everything else, like background music or sound effects, is just an extra layer to enhance that. For example, I love using sound effects, footsteps, a door slamming, the rustling of trees. I grew up listening to radio storytelling, and those subtle sounds really triggered the imagination. So, in a story like *Little Red Riding Hood*, having sound effects like rustling leaves when the wolf is hiding, or a sudden thud, can elevate the whole mood. So yes, music and sound design play a huge role in enhancing the emotional impact of a story.

LM: Yeah, absolutely. For a full application, I think it would make sense to have a simple toggle. So, users could choose if they want background music or sound effects, and maybe an AI model could analyze the mood of the story or specific scenes to decide what kind of music to generate. Or, in the case of sound effects, detect moments like a "shaking sea floor" if that is in the story and add a matching sound clip there.

AD: That would be great, like a little sound bite system. I was actually reflecting on some of the screen recording work we've been doing. We ran into some frustration because our process lacked feedback loops. No one was really checking if things were working well, if the narration quality was good, or if we liked the voice, the color scheme, or the captions. What you're describing is that kind of feedback system but integrated into the storytelling process itself. For example, the AI could ask, "Do you want this with music?" and give a quick preview of what that would sound like. Kind of like the Photo Story software, where you could preview your choices before rendering the whole thing.

LM: Yeah, I really like that idea. I can also imagine a voice preview system where users, especially kids, can choose from different narration options. Male, female, different accents. That sort of customization could make the experience more engaging.

AD: That reminds me, I tried something similar recently. I was working on an English–Afrikaans translation and selected a "South African English" voice. But then I typed something in Afrikaans, expecting the AI voice to read it with the right pronunciation. It completely butchered it, it read Afrikaans with an English accent. So yeah, AI can be quite. unintentionally funny sometimes.

LM: Haha indeed, I know sometimes AI can be quite stupid, to put it that way. Alright, let me go into the closing reflection since we're already at an hour. So just from an overall perspective, what would you consider the strongest and weakest aspects of this project?

AD: Yeah, I think we've touched on it already. The strongest aspect for me is the potential this has to support literacy development in that age group. Like you said, in the absence of a parent or teacher, there's something here that can still assist the learner. But I think the weakness might lie in the variety of the stories generated. And more broadly, my biggest concern is really about creativity and critical thinking, how much of that do we give up when we rely on AI to create things for us? Does it dull our own ability to imagine or think critically?

LM: I definitely understand what you mean. It's similar with coding. AI can help, but there's also a danger if you just let it do everything without understanding. Personally, I only use it when I fully grasp what it's doing, if I know what the code does and why it works, then I'm comfortable using it.

AD: That's a good point. I just read a recent article, Microsoft and Mellon did some early research into how AI affects critical thinking. It's a growing area of concern. I'm curious, would you consider yourself a digital migrant?

LM: Not really. I was born in 2002, so I've kind of grown up with all this. Especially recently, as I've started learning more about how AI works, how large language models like ChatGPT are trained, it's made me very aware of its limits. I know it can be wrong, so I try to use it carefully. For things like emails or feedback, I treat it like a second opinion, like an invisible teammate I can bounce ideas off of. But you have to know how to talk to it and ask the right questions.

AD: Exactly. It's all about balance. Over-reliance is the danger. A lot of people say they use it just for productivity, but I always ask, does it still sound like you? I tell my students the same, if a story doesn't sound like them or feel like their voice, they need to revise it. That's why I advocate for a partnership with AI, not a replacement of the human element.
LM: Yeah, it's a tool, not a substitute.
AD: And I think that over-reliance happens subtly. That's what research is starting to highlight, it creeps in slowly. If people don't have enough experience to recognize when something feels off or isn't written by a human, that's when we start losing something important.
LM: Absolutely. I think human educators are still essential. But AI can help fill gaps, especially when there aren't enough teachers or support in place. If a student doesn't understand something, they can use AI to get help, as long as they've been taught how to use it responsibly.
AD: Yeah, I mean, I just, I met one of my friends who's a principal at one of the, I don't know if you've heard of Bishop's High School?
LM: I don't think so.
AD: Bishops is a private school, one of the wealthiest schools in Cape Town. He's a deputy principal there. He told me they've now decided to ban cell phones completely during school time, not allowed on the premises or in class. It's not really about preventing social media use. From their perspective, it's more about reducing distractions and encouraging students to engage with each other more, rather than being absorbed in their digital worlds. I think that loss of interpersonal engagement really became clear after the pandemic, when students came back and didn't really know how to interact face-to-face anymore. So, there's always this tension, how do we introduce and use technology in a healthy way that doesn't harm the human aspect of who we are?
LM: Yes. Let me also just ask an open question, was there anything else you'd like to add or suggest that we haven't yet covered?
AD: I think I've mentioned this before, but just to highlight again: from an educational perspective, I think it's really important not to encourage passive engagement. Storytelling should involve people actively, both in creating and interacting with the stories. And yes, as we've touched on, AI still carries biases because of how it's trained. So that could limit its usefulness in some contexts, especially if it ends up producing stories with a very Western-centric tone.
LM: Yeah, definitely.
AD: So, the more diverse its training data becomes, the better reach and usability this kind of application will have.
LM: Absolutely. Alright, I think we're about ready to start rounding up. How should I refer to you in my thesis, your full name and role or organization? Do you have a formal title?
AD: Just "Digital Media Coordinator."
LM: And that's with UWC?
AG: Yes, that's right. I'm part of the Center for Innovative Education Communication Technologies, CIECT. It's a bit of a long name, but that's the full title.
LM: Got it!
LM: I'll double-check. And my final question, would you be open to follow-up conversations if I have additional questions later on?
AD: Yes, for sure.
LM: Great! Then I think we're done. Thank you very much for the interview, I really appreciate it. It was super insightful to hear your perspective.
AD: You're welcome. I hope it was useful! Like I said, I wasn't sure at first, AI isn't my field, but this has been a very interesting conversation.
LM: Yeah, as I mentioned, I've interviewed others as well, and I still plan to do one more interview focused more on the technical AI side. But for this one, I really wanted input on the application's purpose, storytelling, literacy, and education.
AD: Can I ask, does your thesis also involve feedback from the target age group? Like, how do children actually experience it?
LM: Not directly. The proof of concept is complete, and while I can keep developing it further, the thesis itself isn't about extending the demo product. It's more reflective, what I made, how I made it, what I learned, and how I'd advise someone else tackling the same question.
AD: Okay, got it.
LM: All right, thank you so much again, and have a great day!
AD: You're welcome. Good luck!
LM: Bye!

Interview transcript – Wouter Grove

Interview with Wouter Grove, manager of the future innovation lab at UWC.

LM: All right then, I think we're ready to start the interview.
WG: Sure.
LM: Let me begin by asking if you had a chance to review the project materials I sent over. Is there anything you'd like me to clarify regarding the project and its objectives?
WG: I looked through it and found it quite interesting. I had a brief look at some of the demos as well. Maybe the immediate question I had is: how do you ensure, for your target audience, children aged 5–8, that the model has the right guardrails? How do you make sure there's no inappropriate content and that the stories are generated at the right level of complexity?
LM: Yeah, no worries. I can say, my prompts are very specific. The story is generated in three parts, beginning, middle, and end. For example, in the beginning prompt, I specify that the model is a skilled storyteller writing a children's story using clear and simple language suitable for that age group. In the code, I also have a hardcoded list of prohibited words, things like crime, adult themes, violence, prejudice, and so on. I asked ChatGPT for a list of words that should never be in a children's story and used that as a baseline. Besides that, the front-end metadata JSON file includes descriptions for each setting and character, which are all designed to be fairy-tale-like and age-appropriate. From what I've seen so far, nothing inappropriate has ever come up in the generated stories.
WG: I figured it would be mainly about careful prompting. That answers my question. I assume that when deploying this to users, it would be in the form of an app or a website with a more refined front-end?
LM: Yeah, definitely. This current front-end is made with Gradio since it's just a simple prototype and that library has nice building blocks. But if I were to imagine it as a full application, say something you could download from the App Store, it would focus more on visuals like big icons and minimal text.
WG: Makes sense. That's all pretty clear and straightforward. One other thing I was curious about, the music generation part. How is the link between the story and the music made? Is it through prompting as well? Do the story elements influence the music prompts?
LM: Currently, the instruments are linked to whichever setting is chosen, and the model uses that information in the music prompts. That said, the music was one of the more challenging parts. The original plan was full background music, but that wasn't really feasible due to the limitations of the available models. So, I ended up using short transition clips instead. The model I use, Facebook's MusicGen-Small, generates three-second clips, and the larger version can generate up to fifteen seconds. But for full background music, you'd need something like a sliding window generation, which takes a lot more time and resources. Considering everything is running locally on my older laptop with a GeForce 2060, I had to be realistic.
WG: Yeah, I feel your pain. That's a lot to handle. What I think would be really fascinating is if you could link this to automatic video generation tools like Sora. That would make for a very marketable product. I can imagine an app where a parent could make a few simple choices and get an auto-generated bedtime story in their language of choice, with visuals. That would be huge, definitely something companies would be interested in buying.
LM: I agree, that kind of visual storytelling would be a great next step. I actually had a similar discussion with Gassant Gamiet, and he also pointed out the value of including visuals. Maybe even having two modes, a full visual version and a simpler narrated one. Let me ask you this as well: based on what you've seen so far, does the core concept of this AI storyteller make sense in an educational or industry context? And what stands out to you the most?
WG: I think it definitely has potential. I can see many applications, especially for children with special needs. What I really like is the randomness in the storytelling, it breaks away from the usual formulaic structure like the Pixar-style stories. In mainstream educational media, kids are often exposed to the same patterns, and I think that can limit creativity. This kind of semi-random storytelling could actually inspire interesting discussions. As a parent, I know my youngest would really enjoy it. It could be a great tool for sparking conversations. Also, since your system is built around modular prompting, you could easily adapt it for targeted stories, like safety at home or other educational themes. The structure stays the same, and you just adjust the prompts. That flexibility gives it a lot of potential.
LM: From a technical perspective, looking at the different parts of the pipeline, story generation, text-to-speech, and music generation, which do you feel is the strongest? And where do you think there might be room for improvement or optimization?
WG: Well, I'm not a hardcore developer, even though I manage software projects, so I won't go into deep technical detail. But for me, the pipeline does make sense. I think a logical next extension would be adding sound effects to the narration. That would add to the interactivity and the immersion of the story. After that,

adding visuals would be the next big step. Another exciting idea would be making the story interactive, letting children influence how the story ends. That would create branching narratives, where a child could make choices that affect the outcome. It's something that could be useful not just in education but also in areas like therapy. I know someone who does play therapy with children, and this kind of interaction could help in understanding how kids make decisions and why.

LM: Yeah, I get what you mean. You could just prompt the model to generate two well-defined story choices and then present them as buttons. The child could choose which one to go with.

WG: Exactly. If you go back to some of the foundational educational frameworks like Bloom's taxonomy and similar models from the e-learning space, then adding this type of interactivity could really strengthen the learning potential. Storytelling on its own is already great, but when you actively involve the child in the decision-making process, it becomes even more valuable for their development.

LM: Do you also think the modular setup I use makes it easier to integrate newer AI models over time? And do you see any value in combining AI components with human-curated elements?

WG: Absolutely. Modular design is really the only way forward right now, given how fast the tech is evolving. The more modular it is, the more adaptable it becomes. And yes, adding a layer where users can influence elements of the story, without affecting safety or structure, could be powerful. You don't want users adjusting the core parameters or safety guardrails, but things like settings or character preferences? That's great. It opens the door to very specific use cases without needing to rewrite everything.

LM: Looking at the educational angle specifically, do you think something like this could help support early literacy, listening comprehension, or imaginative engagement in young children?

WG: Yes, definitely. One thing I'd suggest adding is a kind of evaluation loop. For example, you could generate a few multiple-choice questions based on the story to check for understanding. GenAI models are actually quite good at doing that. If you prompt it right, it could automatically create little comprehension tests. That would add educational value, especially in school settings.

LM: I also wanted to ask your thoughts on the selection mechanism. So first you choose a setting, and then based on that, only the logical character options are shown, for example, you won't get a mermaid in a desert. And then you choose a theme. Do you think this is encouraging enough creativity while still being developmentally appropriate?

WG: I think that part is really well designed. It makes sense and still leaves enough room for creativity without it becoming nonsensical.

LM: You already mentioned visuals as a good future feature. Do you have any other suggestions that might make the app more engaging or educationally effective? I had thought, for instance, that if this were a full application, it might be good to have some voice guidance built in to help guide the child through the process.

WG: That's a good point. You'd want to consider two different usage scenarios: one where the child uses it independently, and another where a parent is guiding them. Designing different onboarding flows for each use case could work well. Parents often run out of new things to talk about or read to their kids, so something like this could be very helpful. And in more professional environments like therapy, it could serve a completely different purpose, but the core idea still works. That said, if you want to go fully in that direction, it would require some conceptual and architectural expansion.

LM: Right, and since you already said you think this could be interesting in a commercial or educational market, especially with added visuals and user interaction. I'm also curious, do you see any potential barriers for schools, publishers, or platforms to adopt a system like this?

WG: I think if you want to bring this into formal school environments, there would definitely be requirements around assessment and reporting. So you'd need tracking, progress monitoring, data logging for how children interact with the stories, things like that. That means having a back-end database and reporting tools. And of course, the moment you're working with children and collecting any kind of data, privacy and data security become major concerns. That would be one of the key things schools or professional users would look into right away. For a system like this, especially if it's used by children independently or casually with parents, it might be useful to include some kind of usage tracking. Parents might want to see how often their children are using it, whether they're finishing stories, and so on. Of course, it's important to think about how you collect and store that data in a privacy-sensitive way. But you've already got access to the data through the models, so it's really just a matter of capturing and presenting it in a helpful way.

LM: Do you think something like an open-source version of this system would appeal to institutions? Or would a hybrid or licensed model make more sense for broader adoption?

WG: That's a tricky one. Open-source can be great, but it assumes a relatively high level of technical skill from the users. From my experience, especially with the teacher training we've done in South Africa, many educators don't have that level of technical capability. So, a fully open-source model might not be ideal. A hybrid approach

could be more realistic. Something where the core is open-source, but there’s a community behind it contributing plugins, features, or adaptations. That would be interesting. But for professional environments like schools or student support services, a commercial or commercial open-source model might be more practical, because those environments often require more guidance and customization. Ideally, you want this to reach as many people as possible but managing that kind of ecosystem takes effort, open-source communities don’t run themselves, even if we like to think they do.
LM: From a business perspective, do you think investing in generative AI is justified here, especially compared to simpler template-based tools? Is AI really the key differentiator?
WG: I think so, yes. Trying to hardcode something like this using basic logic statements like “if this, then that” would take forever and it still wouldn’t feel natural or creative. Generative AI unlocks a lot of new possibilities at a much lower cost, especially now that so many models are open-source. The only real concern is the energy use and overall efficiency of running GenAI models. But in this use case, I’d say it’s worth it. You can build something that feels very polished and professional without needing a full dev team to spend months manually building everything or training new models from scratch. So yeah, I think this is a really good application of GenAI.
LM: And when it comes to safety and quality in the content itself, do you think the current automated validation system is enough? Or would it help to also include human moderation or even another AI model specifically for moderation?
WG: For me, it always comes back to privacy by design. From the start, you need to be thoughtful about things like authentication, are you using something like Google Auth, for instance? What data are you collecting, and why? Ideally, you want to collect as little personal data as possible. You definitely don’t want to store things like children’s photos or identifiable information. On the accessibility side, it’s also worth integrating tools for things like color blindness and general usability from the very beginning. It’s much easier to bake that in during the design phase than to add it later.
LM: That’s a great answer, thank you. I was also specifically thinking about the actual story content, how it’s validated to ensure it’s appropriate.
WG: Ah, yes. For that, I think you’d need some kind of reporting mechanism. Something like a parent-facing feedback form, or even a flagging system if the story seems off. For adult users, that’s relatively easy to implement. But if children are using the app directly, you’d need to think of other mechanisms, maybe something like a child-friendly rating system that asks, “Did anything in the story bother you?” or “Did you like the story?”. You’d need to be careful with how you phrase that, of course. And if you introduce intermediate prompts that change the story in real-time, that adds another layer of complexity. Those prompts would need to be very carefully designed to avoid misuse, I’ve seen some strange things happen with poorly handled prompts.
LM: Yeah, I completely agree. If I were to imagine this as a full application, I’d definitely include a little feedback section. For children, it could be something like a simple 1 to 5-star rating. For adults, it would be more detailed. The data collected could then be cleaned up and used to fine-tune the models over time. Looking a bit more towards the future, when it comes to AI storytelling, do you think things like real-time interactivity or personalized narration would be compelling additions? And what about even more advanced technologies, like integrating AR or VR?
WG: Yeah, I think there are a couple of things that could be interesting in the future. One example is if you could have the model remember previous stories and build on them, creating a kind of story continuity. That could be quite powerful, although it’s a bit complex. As for AR or VR, definitely. That could add real value too. I haven’t really kept up with all the developments in generative video and image tools, but if you can generate visuals, it could be adapted to work with VR or AR formats. The main limitation will probably be processing power. You can compress and optimize a lot, but there’s always a trade-off between visual realism and performance. And you definitely don’t want the story to lag or freeze halfway through a key moment.
LM: Yeah, like imagine a huge cliffhanger and then it just freezes, it would ruin the whole experience.
WG: Yeah, those stories wouldn’t be very popular.
LM: And just as a personal opinion, if future models could generate longer and more coherent background music, do you think that would be a valuable addition? Some people I've asked weren’t sure if they’d like it, while others said it might be nice if you could toggle it on or off.
WG: I think music works best when it’s done well enough that you don’t really notice it, it just enhances the story. You only really notice it when it's missing. If it’s integrated well, reacting to the tone of the story, it could definitely elevate the experience. But it’s not easy. One idea that comes to mind is using sentiment analysis on the story and having the music react dynamically to that, like if a tense scene starts, the music should reflect that tension.
LM: Yeah, that was my idea too. I thought about using something like OpenAI's Whisper to create timestamps throughout the narration. That way, I’d know when the mood changes, such as when a villain appears or a

challenge is introduced, so I could match the music to those shifts with tempo and tone changes.
WG: It's tricky though. A lot of music generation models, like Suno or similar tools, are built for more traditional 3-minute songs with a fixed structure, intro, verse, chorus, and so on. They're not meant to adapt to changing content like a dynamic story. It might be worth checking if there's been any research or development into AI models that generate movie soundtracks, that could be a good fit. It would be fun to explore, though unfortunately it could also disrupt jobs in that industry.
LM: Yeah, that's the tricky balance with AI, right? You want to create something useful and innovative, but you also don't want to completely replace an entire profession or industry with a few prompts.
WG: Exactly. I was just looking at the latest image generation capabilities in ChatGPT, if I were a graphic designer, I'd be feeling the pressure. You can do so much now with a simple prompt.
LM: Yeah, it's crazy what's possible now. Anyway, thank you so much for your time. Just to confirm, I'll be referencing you in my thesis as Wouter Grove, manager of the Future Innovation Lab at UWC.
WG: That's perfect.
LM: And would you be open to any follow-up questions if I need anything further?
WG: Totally fine, happy to help.
LM: Anyway, thanks again and see you soon.
WG: Thanks a lot. Bye!

8.4 Installation manual

Introduction

This appendix outlines the installation process for the AI Storyteller application, developed during the Research Project module. It details required dependencies, setup instructions, and verification steps.

1. Prerequisites

Ensure your system meets the following requirements before proceeding:

- Operating System: Windows 10/11 or Linux
- Storage Space: Minimum 16 GB free
- IDE: Visual Studio Code (recommended)
- Python Version: 3.12
- Docker Desktop: Required for running MeloTTS
- Ollama: Required to run the Llama 3.1 model locally

2. Install Python 3.12

Download Python 3.12 from the official website and follow the installation instructions:
https://www.python.org/downloads/release/python-3120/

After installation, verify Python was installed correctly by executing:
`python --version`

3. Install required Python packages

In the terminal, navigate to the project's root directory and run the following command:
`pip install -r requirements.txt`

4. Install and set up Ollama (Llama 3.1)

Ollama is required to run the Llama 3.1 model locally.
Download Ollama from https://ollama.com.

Then, pull the Llama 3.1 model:
`ollama pull llama3.1`

To test Ollama:
`ollama run llama3.1 "Tell a short story about a hero."`

5. Install Docker and set up MeloTTS

Download Docker Desktop from https://www.docker.com/products/docker-desktop/.
Start Docker and execute:

`docker pull timhagel/melotts-api-server`
`docker run --name melotts-server -p 8888:8080 --gpus=all -e DEFAULT_SPEED=0.9 -e DEFAULT_LANGUAGE=EN -e DEFAULT_SPEAKER_ID=EN-BR timhagel/melotts-api-server`

To verify the server is running, navigate to: http://localhost:8888/convert/tts
Use `docker ps` to confirm container status.

Optional: Testing MeloTTS with Postman

Send a POST request to the endpoint with the following settings:

Header:
Content-Type: application/json

Body:
{

```
 "text": "This is a test of the MeloTTS system."
}
```

6. Install Hugging Face transformers for music generation

This package should install via `requirements.txt`, but you can manually install it using:
`pip install transformers`

7. Launching the AI Storyteller

To run the application, use the following command from the project root:
`python.\app\app.py`

Alternatively, run the script directly from your IDE (e.g., by clicking the “Run” button on app.py).

The application interface will launch in your browser via Gradio.

8.5 User manual

Introduction

This appendix provides the user manual for the AI Storyteller application. It offers step-by-step instructions on using the interface, customizing story content, and accessing the final audio output, including narration and thematic music.

The AI Storyteller is an interactive web application designed to generate personalized children's stories. Each story includes three key elements, setting, characters, and theme, and is narrated with dynamic background music. This manual guides users through the interface and the story generation process.

Navigating the interface

The interface is built using Gradio and provides a structured selection process:

- Setting: Choose the environment for the story (e.g., Magical Forest, Desert Oasis).
- Characters: Select one to three characters relevant to the chosen setting.
- Theme: Define the central theme (e.g., Friendship, Adventure).

These selections shape the narrative and ensure thematic consistency.

Generating a story

- Select story elements: Use the dropdown menus to choose a setting, characters, and a theme.
- Click the button: Press the “Create Your Custom Story!” button.
- Automatic generation: The story is generated in three parts: beginning, middle, and end.
- Validation: If any required element is missing or validation fails, an error will be displayed, and regeneration may occur.

Listening to the story

After generation:

- The full story text appears in a readable text box.
- Narration with music plays via an integrated audio player.
- Press “Play” to listen to the story.

Understanding the outputs

- Story text: The complete story is available for reading within the interface.
- Narration with music: Combines synthesized speech with contextual music, played within the UI.

Credits & licenses

The application uses the following AI models:

- Llama 3.1 (Meta AI) – for story generation.
- MeloTTS – for text-to-speech narration.
- Facebook MusicGen-Small – for music generation.

License information is accessible in the Credits & Licenses section of the user interface.

Troubleshooting

No story is generated: Ensure all required selections are made.
No audio output: Check application access to audio devices and logs for errors.
Mismatched elements: Re-adjust settings or characters to avoid conflicts.

Conclusion

The AI Storyteller offers an enjoyable and customizable way to create narrated children's stories with themed music. This guide helps users explore different combinations of settings, characters, and themes to craft unique and immersive tales.