Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2338~2346
e-ISSN: 2548-6861

Automated Generation of Folklore Short Stories Using T5 Transformer Model

Evangelika Pirade [1]*, I Ketut Gede Darma Putra [2], Desy Purnami Singgih Putri [3]***
* Teknologi Informasi, Universitas Udayana
** Teknologi Informasi, Universitas Udayana
*** Teknologi Informasi, Universitas Udayana
evangelikapirade@student.unud.ac.id[1], ikgdarmaputra@unud.ac.id[2], desysinggihputri@unud.ac.id[3]

Article Info

Article history:

Received 2025-07-30
Revised 2025-09-11
Accepted 2025-09-12

Keyword:

Folklore Generation,
Transformer,
T5,
Reading Interest.

ABSTRACT

High reading interest plays an important role in increasing knowledge and fostering a stronger literacy culture. With the growing access to information and technology, reading interest is also expected to improve through innovative and interactive platforms. However, traditional reading materials often fail to attract younger generations who are more engaged with digital content. To address this challenge, one of the efforts undertaken is the development of a modern platform that provides a collection of short stories enriched with cultural and educational values, tailored to appeal to contemporary readers. This study aims to design and implement a short story generation system using a Transformer-based language model, specifically T5 (Text-to-Text Transfer Transformer). The model is fine-tuned using a curated dataset of folktales from various regions, with the goal of producing relevant, engaging, and coherent narrative texts. The generation process is supported by pre-processing techniques to structure the data into narrative components such as introduction, conflict, climax, and resolution. The generated stories are then evaluated through human evaluation methods, including questionnaires and User Acceptance Testing (UAT), to assess their quality, coherence, engagement, and cultural relevance. This ensures that the system not only produces technically valid texts but also delivers narratives that are meaningful and enjoyable for readers. Ultimately, this study contributes to the promotion of literacy by presenting local wisdom and traditional values from diverse cultures through stories in a more modern, engaging, and accessible format for the younger generation.

I. INTRODUCTION

In language skills, there are several essential components that must be acquired which are speaking, listening, and writing [1]. Out of the four of those components, reading becomes the standard of literacy level dan reading interest. Reading is an essential basic skill that enables individual to acquire knowledge and information from a wide range of sources [1]. However, achieving this also requires a strong reading interest, which should be cultivated from an early age.

Reading interest is the level of desire to enjoy the activity of reading. According to a 2019 survey conducted by the Programme for International Student Assessment (PISA), Indonesia was found to have a very low level of reading interest, ranking 62nd out of 70 countries [2]. National data also reflect similar conditions. The National Library of Indonesia (Perpusnas) reported that the *Tingkat Gemar Membaca* (TGM) score in 2021 was only **59.52** on a 0–100 scale, which still falls in the "moderate" category and indicates that reading has not yet become a strong habit in society [3]. Although the score improved to **63.90** in 2022, the overall trend shows that the reading interest of Indonesians, especially among children and adolescents, remains relatively low [3]. In response to this situation, various solutions can be implemented, one of which is

providing engaging reading materials for children to help foster reading interest from an early age.

One of the interestin reading materials that can be used is folktales that discuss social issues in society and can also introduce cultural diversity to children. Folktales are literary works that describe events occuring in a particular region in acient times [4]. There are many well-known folktales in Indonesia, such as Malin Kundang, Timun Mas, Batu Menangis, and others. Folktales often contain values that can be applied in daily life. Their function serves not only as educational tools—conveying moral messages from the author—but also as a source of entertainment and a reminder of past events or legends that can be passed down from generation to generation.

However, the process of creating folktales manually can be time-consuming and labour-intensive, as it often involves extensive research, such as interviewing local sources and composing the narrative. In addition, the lack of writers who are knowledgeable about folktales presens another challenge on the development such stories [5].

One of the solutions that be applied to handle this problem is to create an automatic system that can generate folktales by implementing Natural Language Processing (NLP). NLP is a branch of artificial intelligence that involves the analysis, understanding, and generaation of human language, enabling interaction with computers through both wirtten and spoken natural language rather than computer-specific language [6]. NLP encompasses a wide range of tasks related to language processing, one of which is Natural Language Generation (NLG). NLG is a subfield of artificial intelligence and computational linguistics that focuses on designing computer systems capable of generating text in human language [7]. Text generation aims to enhace a computer's ability to recognize patterns is text, enabling the development of systems that can comprehend and produce human language.

There has been a lot of research realted to the implementation of Natural Language Generation. Several approaches to text generation utilize deep learning architectures, including Long Short-Term Memory (LSTM) due to its architecture designed to address the vanishing gradient problem in RNNs and its ability to process sequential data such as text. Several studies have implementeed LSTM for text generations tasks, one of which was conducted by Dhall, Vashisth, and Saraswat (2020) [8]. The dataset used in this research was "Alice in Wonderland.txt", which contain 163,780 characters. The preprocessing steps involved converting all characters to lowercase and mapping unique characters into numeral representations. The best result achieved in the study was an accuracy of 71.22%. This study demonstrate that LSTM can help mitigate the vanishing gradient problem commonly encountered in traditional RNNs. Another research by Assabil et al (2023) [9], implemented LSTM to generate Linkedin posts. This research used secondary data obtained from Kaggle, which includes 34.000 Linkedin contents directly uploaded by users. The method used in the study involved two approaches: word-based text generation and character-based text generation, with the best result achieving an accuracy of 94%. However, based on evaluations conducted by four respondents, the generated text still exhibited shortcomings in several aspects, particurarly in terms of grammar. In addition, Santhanam (2020) [10], also implemented LSTM to generate context-based text. The study introduced a method that incorporated context vectors into the training of the text generation model to enhance the semantic meaning of the generated output. The evaluation metric employed was cosine similarity, which measures the semantic alignment between the generated text and the provided context. While the best result achieved an accuracy of 97%, the semantic similarity score based on cosine similarity showed relatively low alignment between the generated text and the training data.

Another study used a dataset consisting of stories with several aspects such as characters, plot, setting, theme, and writing style. These aspects play an important role in shaping a narrative. The research conducted by Hatta Fudholi (2022) [11] aimed to automatically generate stories in Indonesian using skip-thoughts, a technique for representing text with an encoder-decoder model. The encoder used was an RNN based on GRU, while the decoder applied a conditional GRU-based RNN. In this approach, the encoder encodes the input sentences and then passes them to the decoder. The study utilized two datasets: a folklore dataset with 3,872 sentences and a short life-story dataset with 35,897 sentences. The results were obtained from a survey involving five respondents proficient in Indonesian writing. The best results came from experiments using the life-story dataset, which achieved higher qualitative evaluations compared to the folklore dataset that had fewer data. However, both experiments still showed limitations in terms of sentence coherence and overall story unity. Another study used a combination of Markov Chain and Bidirectional GRU (BiGRU) to generate short stories in Indonesian. The aim of this research was to improve model training speed compared to word embedding-based approaches. The model is capable of capturing short-term dependencies using Markov Chain and long-term dependencies through BiGRU. The evaluation was carried out by comparing the developed model with a word-based BiGRU model, a character-based GRU model, as well as models from previous studies. The dataset used in this research consisted of two categories: Indonesian folktales and fairy tales written in Indonesian. The results showed that the Markov Chain-BiGRU model improved training speed by 66.38% compared to other scenarios used in the study. However, the stories generated by this model only showed strengths in two aspects; theme and setting, out of the eleven storytelling aspects used for evaluation [12].

Another study, utilized the Transformer architecture to develop a children's story generator. The research employed

several datasets including the Hugging Face story merge dataset and a dataset provided by fellow researchers. This study cinducted fine-tuning on the GPT-2 model by configuring parameters and processing the dataset accordingly. The study contributes to the future research on story generators by demonstrating the advantages of using Transformer-based methods [13]. Another study conducted by Santoso et al. (2024) [14] applied a similar method by fine-tuning GPT-2 to develop a regional story generator in the Javanese language. This research implemented six scenarios by comparing GPT-2 Medium (Indonesian) and GPT-2 Small (Javanese) models with each respective dataset. The result showed that the model's performaces was relatively poor, and the study did not include any human evaluation to assess the quality or coherence of the generated stories. Another study conducted by Senadeera and Ive (2022) [15] applied a soft prompt tuning approach on the T5 model to generate text that aligns with specific attributes such as positive and negative sentiment. The model was used to generated text conditioned on predefined attributes, particularly sentiment, with a focus on product reviews. The study demonstrated strong model performance and showed that tuning T5 model can help generate more controlled text while preserving the underlying characteristics of the input data.

Based on the studies above, this research adapts several strengths and weaknesses identified from each work. The technical evaluation metric in the form of accuracy showed results that did not align with the desired goal, namely producing stories with good structure and coherence, as well as being consistent with the input expected by users. The method used in this research is T5, which was chosen based on the limitations observed in previous methods such as LSTM and RNN that often struggle to capture long-term dependencies effectively, and GPT models that may generate fluent text but tend to lose control over coherence and alignment with the given input. Furthermore, this choice is supported by the findings of Senadeera's study, which demonstrated that T5 can produce more controlled results while maintaining the characteristics of the input provided.

In addition, T5 is more commonly used for tasks involving textual data, such as simplification, classification, and others. In previous studies that serve as references for this research, T5 was applied to text generation tasks to produce text based on the desired sentiment and achieved promising results. Therefore, in this study, T5 is employed to generate short stories.

II. METHOD

This research, implemented the steps of text generation by Reiter and Dale (1995) [7] which are content planning, text structuring, and surface realization. The methodological steps undertaken in this study are presented in Figure 1.

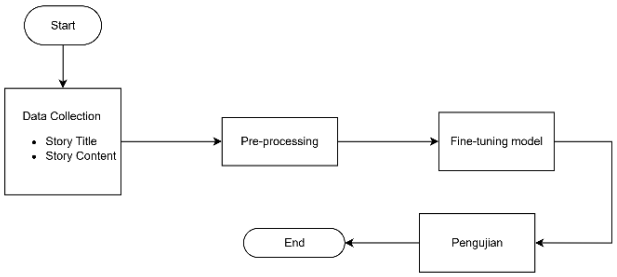

Figure 1 Research Flow

Figure 1 represents the research flow outlining each step from data collection to evaluation. The research flow begins with collection short stories categorized as folktales from both online sources and printed materials. The collected data then undergoes preprocessing, which includes cleaning the text by removing punctuation marks and unnecessary elements. This step also involves standardizing the story structure into four segments: introduction, conflict, climax, and resolution. After cleaning, the data is tokenized using the tokenizer from the T5 model. The tokenized dataset is then used to fine-tune the T5 model. The final stage of this research involves human evaluation.

A. Content Planning

In the content planning stage, data collection is carried out alongside prompt engineeringm which is used to design standardized prompt for model training. Additionally, a query rewriting process is performed to transform user input into a story representation that can be effectively utilized by the system.

1) *Data Collection*: This research utilizes two types of data obtained from various sources, namely primary data and secondary data. Primary data refers to data collected directly during the research process and is used to address and resolve the research problem. The primary data in this study consists of short stories from various countries. Secondary data, on the other hand, is obtained from external sources such as books, papers, articles, and others. The secondary data serves as a theoretical reference and supporting foundation in both conducting the research and preparing the report. In the data collection process, two methods were employed, namely manual scraping and scanning. Manual scraping was conducted by extracting relevant content from online sources, while scanning was applied to printed sources. To ensure data quality, the collected texts were subsequently reviewed, cleaned, and validated to remove duplicates, irrelevant content, or errors resulting from the extraction process. This step was carried out to guarantee that the dataset is consistent, reliable, and ready for further preprocessing and analysis. The data collection process began with scraping folklore stories by copying the titles and contents from online sources and scanning printed sources. This approach was necessary because the folklore stories originated from various sources and lacked a standardized

structure, making it difficult to extract the data automatically. Afterward, the collected data was compiled into an Excel file to facilitate further implementation.

2) *Prompt Engineering*: Prompt engineering is utilized to form the intial representation of the user's input. This method functions to design prompt or instructions so that the model can understand the intended direction of the story, even when the input is unstructured or freely written by the user. This method is applied during the model training phase by preparing standardized prompts for each data entry. These prompts include clear and consists instruction such as "Make a story from this storyline:..", followed by a brief summarization of each part of the story in the dataset.

3) *Query Rewriting*: This stage is used to interpret user inputs that are free-form or unstructured into explicit narrative representations. These representations take the form of structured prompts, which are then user to train the model. This process is applied during evaluation with real users, where the user's intial input is transformed into a predefined prompt structure suitable for model training. This method utilizes the OpenAI API with the Chat-GPT 4o model, which is capable of understanding natural language context and semantically rewriting inputs to better align with the model's requirements. The following Table is an example of how a query is transformed.

TABEL I
TRANSFORMED QUERY

Input User	Rewrited Query
Aku ingin cerita tentang dua kakak beradik yang tersesat di hutan terlarang dan harus memecahkan teka-teki alam untuk bisa pulang.	Buatlah cerita dengan alur seperti berikut: Di sebuah desa kecil dekat hutan misterius, hiduplah dua saudara petualang bernama Arin dan Lila. Meski sudah diperingatkan, mereka tetap masuk ke hutan terlarang dan menghadapi tantangan dari alam, seperti teka-teki angin dan hewan yang bisa bicara. Dengan kerja sama dan kecerdikan, mereka berhasil menyelesaikan semua tantangan dan belajar tentang keberanian serta keharmonisan dengan alam. Setelah kembali ke desa dengan selamat, mereka membagikan kisahnya untuk menginspirasi orang lain tentang keajaiban alam.

B. Text Structuring

In this stage, the elements of the story are organized to form a systematic narrative structure. This step involves data preprocessing that includes data cleaning, data translation, and dividing the story into four main sections: the opening, the conflict, the climax, and the resolution.

1) *Data pre-processing*: Pre-processing is a crucial stage that involves preparing the data before it is used in model training. This stage includes data cleaning which includes removing unnecessary punctuation, special characters, extra spaces, and other irrelevant text attributes. Then, each story is divided into four parts; the opening, conflict, climax, and resolution. This is intended to ensure that the generated stories remain coherent, engaging, and aligned with traditional narrative structures. After segmentation, the contents of the story are translated from Indonesian into English. This step is necessary because the base model used in this study was pre-trained on English corpus, making it more effective in processing and understanding training data in the same language. Once the data is translated, tags are added to mark the beginning of each story part, and the sections are then recombined. This allows the model to recognize the overall story structure more effectively. Finally, tokenization was performed using the T5Tokenizer from the T5 model used in this study. Prior to tokenization, the dataset was first converted into the Hugging Face Dataset format to endure compability with the training pipeline. The dataset was then split into training and testing sets with a ratio of 70:30

C. Surface Realization

Surface realization refers to the final output of the story that incorporated grammatical structures and linguistic coherence. In this study, this stage begins with model training and continues through to the human evaluation of the generated stories. The goal is to produce narratives that are not only relevant in content but also grammatically correct and natural in language.

1) *Model Training*: The training process refers to the stage in which the prepared model is fine-tuned to generate folk tales. The model is trained using the preprocessed and structured dataset to ensure that the outputs are relevant, coherent, and aligned with traditional narrative forms.

TABEL II
MODEL TRAINING PARAMETERS

Parameter	Value
Model	t5-small
Epoch	20
Batch Size	4
Learning Rate	5e-5

The training process for generating the story content begins with initializing the pre-trained model used in this research, which is t5-small. This model is chosen due to its efficient architecture and proven capability in handling natural language processing tasks, particularly text generation. The T5-small model is designed with a transformer-based encoder–decoder architecture, consisting of approximately 60 million parameters, 6 encoder and decoder layers, and 8 attention heads per layer [16]. This model can be fine-tuned for specific tasks such as text generation. The model was chosen due to its relatively small size while still being capable of delivering strong performance in NLP tasks. In this study, T5-small was fine-tuned using a folklore dataset that had been translated into English and structured according to narrative components

(introduction, conflict, climax, and resolution). The hyperparameters—which play a crucial role in the training process—are set. Some key hyperparameters used in this study, as shown in Table II, include:

- Epoch: 20, indicating the number of complete passes through the training dataset.
- Batch size: 4, referring to the number of samples processed through the training dataset.
- Learning rate: 5e-5, which determines how much the model's weights are adjusted during training.

After all the parameters are configures, the training is carried out using a dataset of folktales that has been pre-processed and structured into narrative elements. The model is trained until it converges or reaches stable performaces. The goal of this training process is to enable the model to generate coherent, relevant, and narratively structured folk tales that follow the traditional storytelling flow.

2) *Evaluation*: The evaluation applied in this study is human evaluation. In the context of automatic text generation, human evaluation refers to the process in which individuals or groups of human assessors evaluate the quality of texts generated by the system evaluation is essential to obtain feedback on how well the generated text meets human standards in terms of relevance, coherence, creativity, and other qualitative aspects [17]. Human evaluation helps researchers verify the quality of outputs from story generation systems, which cannot always be reliably captured by automatic metrics. This evaluation method was chosen in consideration of previous studies, which showed limitations in technical evaluation metrics. Such metrics often yield results that differ from direct human assessment, whereas the ultimate objective of story generation is to produce narratives with well-structured, coherent, and meaningful content that align with human expectations.

In this study, human evaluation was carried out through two approaches: questionnaires and User Acceptance Testing (UAT). The questionnaire was distributed to respondents to assess narrative quality based on predefined criteria such as structure, coherence, and creativity. Meanwhile, UAT was conducted to evaluate the usability and user satisfaction of the system, ensuring that the generated stories not only meet narrative quality standards but are also engaging and accessible to the intended audience.

The human evaluation process involved two categories of respondents: university student majoring in Indonesian language and literature, and children aged 8 to 12 accompanied by their parents. This study used three evaluation criteria (1) relevance to the input, (2) coherence and story structure (used by the university student respondents), and (3) story appropriateness for children (used by the child respondents). The evaluation used a 5-point Likert Scale. The process was conducted by having users interact with the application and fill out a prepared evaluation form. Table III to V are the questions used in human evaluations.

TABEL III
QUESTIONS FOR THE ASPECT OF RELEVANCE TO THE INPUT

No.	Questions
1.	Apakah cerita yang dihasilkan sesuai dengan tema yang diinginkan?
2.	Apakah latar cerita mencerminkan konteks yang diminta?
3.	Seberapa baik cerita ini dalam memenuhi ekspektasi dari input pengguna?

TABEL IV
QUESTIONS FOR THE ASPECT OF COHERENCE AND STORY STRUCTURE

No.	Questions
1.	Apakah pembuka cerita cukup jelas dalam memperkenalkan tokoh dan latar cerita?
2.	Apakah alur cerita cukup berkembang secara logis dari pembuka, konflik, hingga puncak cerita?
3.	Apakah penutup cerita memberikan kesimpulan yang memadai dan menyatukan keseluruhan cerita?
4.	Apakah cerita yang dihasilkan sudah mencakup unsur-unsur cerita secara umum (tokoh, latar, konflik, alur, dan penutup)?

TABEL V
QUESTIONS FOR THE ASPECT OF STORY APPROPRIATNESS FOR CHILDREN

No.	Questions
1.	Apakah cerita yang dihasilkan sesuai dengan tema yang dimasukkan?
2.	Apakah kosakata dalam cerita sesuai dengan tingkat pemahaman anak?
3.	Apakah anak menunjukkan ketertarikan saat membaca cerita?
4.	Apakah cerita mengandung pesan moral atau pembelajaran sederhana?

The questionnaire data, which uses weighted Likert scale scores, is then processed by multiplying each weight by the total score obtained. The results are interpreted in the form of percentages. The formula used to calculate the percentage of the human evaluation results is based on Equations (1).

$$p = \frac{f}{N} x\ 100\% \quad (1)$$

Note:
f = frequency of responses obtained
N = total number of respondents

This formula is used to calculate the mean score of the responses provided by the participants, thereby offering a general overview of the level of agreement regarding the assessed aspects as a whole. The interpretation of the resulting values in the form of percentages is then categorized into several predefined categories.The result is

then interpreted in the form of a percentage, is then categorized into several levels.

TABEL VI
ASSESMENT RESULT CATEGORIES

Percentage	Categories
0%-20%	Sangat Kurang Baik
21%-40%	Kurang Baik
41%-60%	Cukup Baik
61%-80%	Baik
81%-100%	Sangat Baik

User Acceptance Testing (UAT) in this study was employed to assess the overall feasibility of the system, ensure that the features within the application function properly, and evaluate whether the system meets user needs and expectations. The following presents the evaluation scenarios conducted through UAT.

TABEL VII
ASSESMENT RESULT CATEGORIES

No.	Fitur yang Diuji	Skenario Pengujian	Hasil yang Diharapkan	Status (Lulus /Gagal)
1.	Welcome page	Pengguna menekan tombol “Mulai Membaca”	Sistem menampilkan halaman input	
2.	Input prompt	Pengguna memasukkan prompt cerita	Sistem menerima input tanpa error	
3.	Generate cerita	Pengguna menekan tombol “Generate”	Cerita (judul dan isi) ditampilkan berdasarkan input	
4.	Struktur cerita muncul	Cerita tampil lengkap: [PEMBUKA] [KONFLIK] [PUNCAK] [PENUTUP]	Struktur naratif tampil dengan jelas	
5.	Input prompt ulang	Pengguna menekan tombol back	Kembali ke halaman input	

III. RESULT AND DISCUSSION

A. Content Planning Results

The result of content planning involves data collection and prompt creation. The generated prompts are used to train the model so that it can produce stories based on the context provided in the prompts.

1) *Data Collection Result*: Data collection is the first stage carried out in this study. The data was gathered by copying the content and titles of the stories. The results of the data collection are shown in the Table VIII.

TABEL VIII
DATA COLLECTION RESULT

Judul	Isi
Batu Menangis	Alkisah, di sebuah desa terpencil hiduplah seorang janda tua dengan seorang putrinya yang cantik jelita bernama Darmi. Mereka tinggal di sebuah gubuk yang terletak di ujung desa. Darmi memang cantik, parasnya indah menawan. Namun, tingkah lakunya sangatlah tidak cantik dan sifatnya sangatlah tidak menarik. Setiap hari Darmi selalu bersolek di kamarnya. Ia tidak pernah mau membantu ibunya sedikit pun membereskan isi rumah. Kamarnya selalu berantakan. Darmi tidak peduli akan hal itu, ia hanya peduli pada wajahnya yang cantik jelita tiada terkira haruslah selalu tampil sempurna. Ibunya Darmi yang sudah tua, setiap hari selalu bekerja keras demi mendapatkan uang. Apapun jenis pekerjaannya, selama itu halal, akan ia kerjakan. Semua itu ia lakukan hanya untuk memenuhi kebutuhan hidupnya dan kebutuhan Darmi, anak semata wayangnya. Ibunya Darmi juga kerap diperlakukan seperti pembantu. Setiap ditanya siapa yang berjalan di belakangmu, ia selalu menjawab bahwa ibunya adalah budaknya. Mendengar hal itu terus menerus, Ibu Darmi merasa sakit hati hingga berdo'a. Secara perlahan Darmi berubah menjadi batu. Ia terus menangis dan memohon kepada ibunya. Namun, semua sudah terlambat. Kini tubuhnya berubah menjadi batu yang terus mengeluarkan air mata.

The collected data consists of story titles and contents. The gathered data is saved in .xlsx format. The stories were gathered from several books containing folktales from various countries, obtained from both online and printed resources. A total fo 923 stories were collected, each including both the title and content.

2) *Prompt Engineering Result*: The result of prompt engineering is a set of prompts designed to guide the model in generating new stories. Table IX contained the outcomes of the prompt engineering process.

TABEL IX
PROMPT MAKING RESULT

Isi Cerita	Prompt
[PEMBUKA] Once upon a time, in a kingdom on the island of Sumatra, there lived a princess named Pukes. Pukes was fond of a prince. [KONFLIK] One day, Pukes and the Prince got married. After the	Make a story with this storyline: Once upon a time, in a kingdom on the island of Sumatra, there lived a princess named Pukes who was fond of a prince. One day, Pukes and the Prince got married, and as they left,

wedding, Pukes wanted to follow the Prince and bid farewell to her parents, who advised her not to look back during her journey. [PUNCAK] However, during the journey, Pukes forgot the message and looked back. As a result, her body turned to stone, and she regretted not listening to her parents' advice. [PENUTUP] Finally, the stone of Putri Pukes still exists to this day, and a lake formed around the stone, known as Lake Laut Kawar.	her parents advised her not to look back during her journey. However, Pukes forgot the message and looked back, causing her body to turn to stone, and she regretted not listening to her parents' advice. Finally, the stone of Putri Pukes still exists to this day, surrounded by a lake known as Lake Laut Kawar.

B. Text Structuring Results

The result of text structuring is a story that has been cleaned and systematically organized based in its content. In addition, the story is transformed into a format that can be processed by the model.

1) *Data Pre-processing Results*: Pre-processing consists of several steps, namely data cleaning, splitting the story into four parts, and adding tags to each section. Table X contaned the outcomes of data cleaning.

TABEL X
DATA PRE-PROCESSING RESULT

Sebelum	Sesudah
Alkisah, di sebuah desa terpencil hiduplah seorang janda tua dengan seorang putrinya yang cantik jelita bernama Darmi. Mereka tinggal di sebuah gubuk yang terletak di ujung desa. Darmi memang cantik, parasnya indah menawan. Namun, tingkah lakunya sangatlah tidak cantik dan sifatnya sangatlah tidak menarik. Setiap hari Darmi selalu bersolek di kamarnya. Ia tidak pernah mau membantu ibunya sedikit pun membereskan isi rumah. Kamarnya selalu berantakan. Darmi tidak peduli akan hal itu, ia hanya peduli pada wajahnya yang cantik jelita tiada terkira haruslah selalu tampil sempurna. Ibunya Darmi yang sudah tua, setiap hari selalu bekerja keras demi mendapatkan uang. Apapun jenis pekerjaannya, selama itu halal, akan ia kerjakan. Semua itu ia lakukan hanya untuk memenuhi kebutuhan hidupnya dan kebutuhan Darmi, anak semata wayangnya. Ibunya Darmi juga kerap diperlakukan seperti pembantu. Setiap ditanya siapa yang berjalan di belakangmu, ia selalu menjawab bahwa ibunya adalah budaknya. Mendengar hal itu terus menerus, Ibu Darmi merasa sakit hati hingga berdo'a. Secara perlahan Darmi berubah menjadi batu. Ia terus menangis dan memohon kepada ibunya. Namun, semua sudah terlambat. Kini tubuhnya berubah menjadi batu yang terus mengeluarkan air mata.	[PEMBUKA] Once upon a time, in a village on the island of Kalimantan, there lived an old widow with her beautiful daughter named Darmi. Despite her lovely appearance, Darmi had a bad attitude and was reluctant to help her mother. [KONFLIK] One day, Darmi once again treated her mother like a servant and arrogantly referred to her as a slave in front of others. Her mother, who was already tired and heartbroken, could only be patient, even though her heart was wounded. [PUNCAK] However, the mother's patience finally ran out, and she prayed for Darmi to learn a lesson for her behavior. Suddenly, Darmi's body began to harden and slowly turned into stone. [PENUTUP] In the end, despite her tears and pleas for forgiveness, Darmi could not escape her fate. Her body had turned to stone, and to this day, the stone continues to shed tears as a symbol of eternal regret.

The dataset then split into training set and a validation set using a 70:30 ratio, resulting in 646 training samples and 277 testing samples. Following the data split, the dataset was tokenized using the T5Tokenizer corresponding to the T5 model used in the experiment.

C. Surface Realization Results

The surface realization stage produced both the model training analysis and the generated stories based on the T5 model. This process aimed to ensure that the output stories adhered to appropiate grammatical structures and narrative coherence. The T5 model was fine-tuned using the preprossed dataset, which included segmented story parts labels with narrative tags to guide the model during generation.

The training was configured with a batch size of 8, and evaluation strategy based on epochs, with a total of 20 epochs, and a learning rate of 5e-5. During the first three epochs, training loss values were not recorded, likely due to the training configuration not enabling logging during the early stages.

Starting from the fourth epoch through the tenth epoch, the training loss values remained constant. This indicates that the model experienced no significant improvement in its learning process during this period. However, the validation loss continued to decrease, which suggests that the model's performance on the validation data was improving, despite the stagnation in training loss. This pattern implies that while the model's training behavior appeared static, it was still generalizing better to unseen data, possibly due to regularization effects or improvements in handling the validation set.

Below is one example of a generated story produced by the system based on a user's input prompt.

TABEL XI
GENERATED STORY RESULTS

Input User	Generated Stories
Ceritakan kisah seorang anak laki-laki miskin yang menemukan benda ajaib di hutan.	[PEMBUKA] Pada suatu ketika, di sebuah desa kecil yang terletak di tepi hutan yang luas dan misterius, hiduplah seorang anak laki-laki miskin bernama Arif. Meskipun awal kehidupannya sangat sederhana, Arif dikenal karena hatinya yang baik dan semangat petualangnya. [KONFLIK] Suatu hari, saat menjelajahi kedalaman hutan, Arif menemukan sebuah lampu kuno yang berkilauan setengah terkubur di tanah. Dengan terkejut, muncul seorang jin mistis yang menawarkan tiga permintaan kepada Arif. [PUNCAK] Namun, dengan kebijaksanaan yang melampaui usianya, Arif menjadi pahlawan yang dicintai, dan lampu ajaib itu dikembalikan ke hutan, menunggu jiwa layak berikutnya untuk menemukan rahasianya.
Aku ingin cerita tentang seorang putri kerajaan yang berani melawan penyihir jahat demi menyelamatkan rakyatnya.	[PEMBUKA] Dahulu kala, di sebuah kerajaan yang jauh terletak di antara hutan lebat dan pegunungan yang menjulang tinggi, hiduplah seorang putri pemberani bernama Elara. Kerajaan tersebut makmur di bawah pemerintahan ayahnya yang bijaksana, Raja Alden. [KONFLIK] Namun, kedamaian hancur ketika seorang penyihir jahat bernama Morgath turun ke tanah itu, melancarkan mantra gelap yang membawa keputusasaan dan penderitaan kepada rakyat. [PUNCAK] Namun, dalam pertempuran klimaks, Elara dengan berani melawan penyihir tersebut, menggunakan keberanian dan kekuatan cinta untuk mematahkan sihir gelap Morgath. [PENUTUP] Akhirnya, setelah penyihir tersebut dikalahkan, kerajaan dipulihkan ke kejayaannya yang dulu dan menjadi pahlawan. Dia berjuang dengan kebaikan dan kekuatan. Dia memerintah dengan kekuatan cinta untuk rakyat kerajaan."

The story generation results presented in Table XI demonstrate that the system is capable of producing complete narratives based on user input. The generated stories are structured into four main narrative components, namely introduction, conflict, climax, and resolution. This indicates that the model is able to follow the expected classical storytelling pattern. Overall, the findings show that the processes of query rewriting and story generation can be effectively applied to produce coherent, logical, and complete narratives that align with the user's initial request.

To asses the quality of the generated stories, a human evaluation was conducted based on the three main aspects. The evaluation involved two groups of respondents. Each group rated a series of questions in a Likert Scale ranging from 1 to 5. The evaluation results are summarized in Table XII.

TABEL XII
AVERAGE HUMAN EVALUATION RESULTS

No	Evaluation Aspect	Question Items	Percentage
1.	Relevance to User Input	A1, A2, A3	90.67%
2.	Story Coherence and Structure	B1, B2, B3, B4	87.50%
3.	Appropriateness for Children	C1, C2, C3, C4	89.00%

The human evaluation results further validate the effectiveness of this approach. As shown in Table 4.10, the model achieved a 90.67% score for Relevance to User Input, reflecting a high ability to respond appropriately to prompts. In the Story Coherence and Structure category, it scored 87.50%, indicating good narrative flow and logical consistency. For the Appropriateness for Children aspect, the model earned a score of 89.00%, confirming that the stories were suitable and understandable for the target audience. For UAT, the results showed that most features operated as expected.

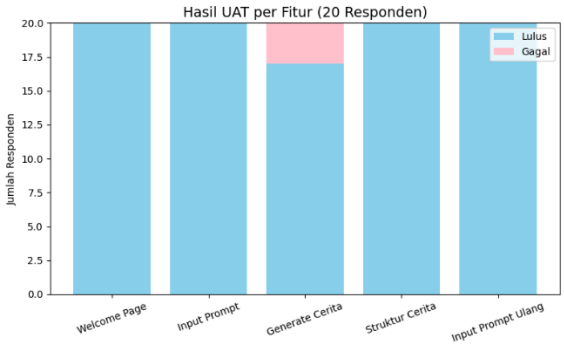

Figure 2 UAT Results

The Welcome Page, Input Prompt, Story Structure, and Re-input Prompt features achieved a 100% success rate without any failures. Meanwhile, in the Generate Story feature, 3 respondents encountered failures due to timeouts, while 17 respondents successfully obtained stories according to their inputs. Overall, from 100 testing scenarios conducted (5 features × 20 participants), 97 scenarios were successful and only 3 scenarios failed. This indicates that the system's success rate falls into the "very high" category, suggesting

that the application has met user needs in generating stories according to the tested scenarios. The failures observed were primarily due to technical factors (timeouts), indicating that future improvements should focus on optimizing system performance to ensure more consistent story generation.

In conclusion, the combination of proper data preparation, structured story design, and targeted prompt engineering significantly improves the quality, coherence, and relevance of short story generation using a T5-based language model.

IV. CONCLUSION

This study successfully implemented a series of steps to build a model capable of generating short stories based on given prompts. The process began with data collection from various sources, followed by data cleaning and structuring to ensure the information was well-organized and suitable for training. Through prompt engineering, meaningful and context-aware prompts were designed to guide the model in generating coherent and relevant stories. The processing phase, which included tagging and dividing the stories into four structured parts, further enhanced the quality and consistency of the dataset. The results of the human evaluation further affirm the effectiveness of the proposed approach. The model demonstrated strong performance in generating contextually appropriate and linguistically sound stories, with the highest average score of 4.53 in the aspect of relevance to user input. The story coherence and structure aspect also received a positive average score of 4.38, while the suitability of stories for children achieved 4.45, indicating that the generated stories were well received by both the target readers and language experts.

Nevertheless, several limitations remain. First, the model is highly dependent on the quality and specificity of the prompts used, meaning that vague or ambiguous prompts may lead to less coherent results. Second, the lack of robust technical evaluation metrics makes it difficult to fully assess story quality before implementation, leaving the evaluation primarily reliant on human feedback. Third, due to these challenges, the model still depends on the default hyperparameters of T5, limiting opportunities for deeper optimization and fine-tuning. Despite these limitations, the system provides a significant improvement over traditional reading materials. Unlike static and one-size-fits-all content, this approach allows for automatic story generation that can be tailored to user preferences, ensuring more engaging and personalized reading experiences. By integrating folklore and cultural narratives into an interactive digital format, the system not only preserves traditional values but also reintroduces them to younger generations in a modern and accessible way. This highlights the potential of AI-assisted storytelling as an innovative tool to foster literacy, creativity, and cultural appreciation in today's digital era.

REFERENCES

[1] S. Kasiyun, 'Upaya Meningkatkan Minat Baca Sebagai Sarana Untuk Mencerdaskan Bangsa', 2015. [Online]. Available: http://journal.unesa.ac.id/index.php/jpi

[2] H. Retno, 'Miris, Minat Baca di Indonesia Menurut UNESCO Hanya 0.001 Persen', Bandung, May 17, 2021.

[3] T. Nurrahim, 'Orang Indonesia Makin Gemar Baca', 2023. Accessed: Sep. 11, 2025. [Online]. Available: Orang Indonesia Makin Gemar Baca

[4] M. Shofiyullah, P. Bahasa, D. Sastra, I. Fakultas Bahasa, and D. Seni, *Pesan Moral Dalam Kumpulan Cerita Rakyat Nusantara Karya Yustitia Angelia Sebagai Bahan Ajar Pembelajaran Mengidentifikasi Nilai dan Isi Cerita*. 2020.

[5] S. Chakraborty, 'A Study of Folklore-Publication', 2019. [Online]. Available: https://www.researchgate.net/publication/372492438

[6] H. A. Parhusip, 'Machine Learning dengan Pyhton'.

[7] E. Reiter and R. Dale, 'Building Applied Natural Language Generation Systems', Cambridge University Press, 1995.

[8] I. Dhall, S. Vashisth, and S. Saraswat, 'Text Generation Using Long Short-Term Memory Networks', in *Lecture Notes in Networks and Systems*, vol. 106, Springer, 2020, pp. 649–657. doi: 10.1007/978-981-15-2329-8_66.

[9] M. R. Assabil, N. Yusliani, and A. Darmawahyuni, 'Text Generation using Long Short-Term Memory to Generate a LinkedIn Post', *Sriwijaya Journal of Informatic and Applications*, vol. 4, no. 2, pp. 57–68, 2023, [Online]. Available: http://sjia.ejournal.unsri.ac.id

[10] S. Santhanam, 'Context based Text-generation using LSTM networks', Apr. 2020, [Online]. Available: http://arxiv.org/abs/2005.00048

[11] D. Hatta Fudholi, 'Story Generator Bahasa Indonesia dengan Skip-Thoughts', 2022.

[12] C. Angieta Winata, H. Santoso, and I. Wasito, 'Word-Level Indonesian Story Generator With Markov Chain And Bidirectional GRU', vol. 10, no. 4, pp. 14–26, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[13] B. Hettige, V. Bandara, A. Sanja, R. Bandara, and H. Sanja, 'Sibil AI: Children Story Generator in Sinhala Using Transformers', 2022. [Online]. Available: https://www.researchgate.net/publication/364787507

[14] K. Q. Santoso, K. Perjuangan Mustadl'afin, L. Andriyanto, and I. Ardiyanto, 'Generator Kisah Daerah Berbasis Bahasa Jawa dengan Finetuned GPT-2', 2024.

[15] D. C. Senadeera and J. Ive, 'Controlled Text Generation using T5 based Encoder-Decoder Soft Prompt Tuning and Analysis of the Utility of Generated Text in AI', Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.02924

[16] google-research, 'text-to-text-transfer-transformer'. Accessed: Jan. 06, 2025. [Online]. Available: https://github.com/google-research/text-to-text-transfer-transformer

[17] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, 'Human evaluation of automatically generated text: Current trends and best practice guidelines', *Comput Speech Lang*, vol. 67, May 2021, doi: 10.1016/j.csl.2020.101151.