Text-to-Text Automatic Story Generation: A Survey

Yuan Ma[1], Richard Susilo[1], Patrik Haslum[1], Hanna Suominen[1,2,3]
[1] School of Computing, The Australian National University, Canberra, ACT, Australia
[2] School of Medicine & Psychology, The Australian National University
[3] Department of Computing, University of Turku, Turku, Finland
{yuan.ma, RichardReynaldo.WironotoSusilo, patrik.haslum, hanna.suominen}@anu.edu.au

Abstract

Automatic story generation aims to produce coherent, engaging, and contextually consistent narratives with minimal or no human involvement, thereby advancing research in computational creativity and applications in human language technologies. The emergence of large language models has progressed the task, enabling systems to generate multi-thousand-word stories under diverse constraints. Despite these advances, maintaining narrative coherence, character consistency, storyline diversity, and plot controllability in generating stories is still challenging. In this survey, we conduct a systematic review of research published over the past four years to examine the major trends and key limitations in story generation methods, model architectures, datasets, and evaluation methodologies. Based on this analysis of 57 included papers, we propose developing new evaluation metrics and creating more suitable datasets, together with ongoing improvement of narrative coherence and consistency, as well as their exploration in practical applications of story generation, as actions to support continued progress in automatic story generation.

1 Introduction

Stories are a significant part of our lives. They accompany us as we grow, shaping our cognitive development and understanding of the world (Merrill and Fivush, 2016). Stories also serve as an essential medium for communication, helping bridge the gap between the writer's knowledge and the listener (Suzuki et al., 2018). Writing compelling stories is a deeply creative process that has long been considered difficult to emulate with Artificial Intelligence (AI) (Anantrasirichai and Bull, 2022).

With advances in technology, automatic story generation has gained increasing attention for story writing, leading to a range of models designed and developed to achieve this task (Alhussain and Azmi, 2021; Fang et al., 2023; Teleki et al., 2025). Automatic story generation involves selecting a sequence of events or actions that meet specific criteria and can be presented as a coherent narrative within a story world featuring distinct characters (Li et al., 2013). Recent surveys (Alhussain and Azmi, 2021; Fang et al., 2023) indicate that traditional approaches, such as planning-based methods, once dominated story generation research. Over the past four years, however, rapid advancements in Large Language Models (LLMs) have driven substantial progress: recent LLM-based methods demonstrate in evaluation studies (Gómez-Rodríguez and Williams, 2023; Tian et al., 2024) and surveys (Teleki et al., 2025) the capability to produce higher-quality stories that are substantially longer, and efforts to leverage LLMs for evaluation have brought automatic assessment methods closer to human judgments.

This survey aims to systematically review the evolution of methods in story generation, identify key challenges, and outline promising directions for future research. The work most closely related to ours is the survey by Teleki et al. (2025), which focuses specifically on the recent use of LLMs for story generation, as well as the datasets and evaluation metrics employed in these approaches. In contrast, our survey takes a broader perspective, without restricting the scope to LLM-based methods, covering general story generation research from 2021 to 2025, including studies introducing newly proposed datasets and evaluation metrics. Our research questions are as follows:

- What Natural Language Processing (NLP) techniques have been employed in automatic text-to-text story generation over the past four years?
- What challenges remain for current approaches to automatic story generation?

In this work, we (1) collect a set of 57 peer-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 514–527

reviewed story generation papers published within the past four years and categorize them into three methodological groups; (2) analyze and quantify five datasets and nine evaluation metrics most commonly adopted across included studies; (3) highlight two enduring challenges that existing methods continue to face; and (4) propose three future research directions together to guide further advancements in this area.

2 Methods

This paper surveys research published between January 2021 and April 2025 on automatic text-to-text English story generation using NLP techniques. The Association for Computational Linguistics (ACL) Anthology is used as the primary database, and a keyword search is applied to paper titles and abstracts (Table 1).

To ensure comprehensive coverage of the most relevant and impactful works, an additional search is conducted across five leading linguistic conferences website with the highest h-index scores: the Annual Meeting of the ACL, the Conference on Empirical Methods in NLP (EMNLP), the Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT), Transactions of the ACL (TACL), and the International Conference on Computational Linguistics (COLING). Owing to constraints inherent to the conference websites, searches are restricted to keyword queries applied to paper titles (Table 1).

In this survey, we limit our focus to text-to-text story generation, excluding related NLP tasks like image-to-story generation. We have also excluded studies that center on story generation in non-English languages. Two reviewers jointly screen the titles and abstracts to identify relevant papers, while one reviewer summarizes and analyzes the selected works. In total, 57 articles are identified from searches across the five targeted conferences and the ACL Anthology database.

3 LLM capacity in Story Generation

LLMs have achieved exceptional performance across various natural language generation tasks, including machine translation (Zhu et al., 2024) and text summarization (Zhang et al., 2024). Part of the included studies investigates the use of LLMs for story generation, with a particular emphasis on evaluating their narrative and storytelling capabilities rather than proposing new generation techniques.

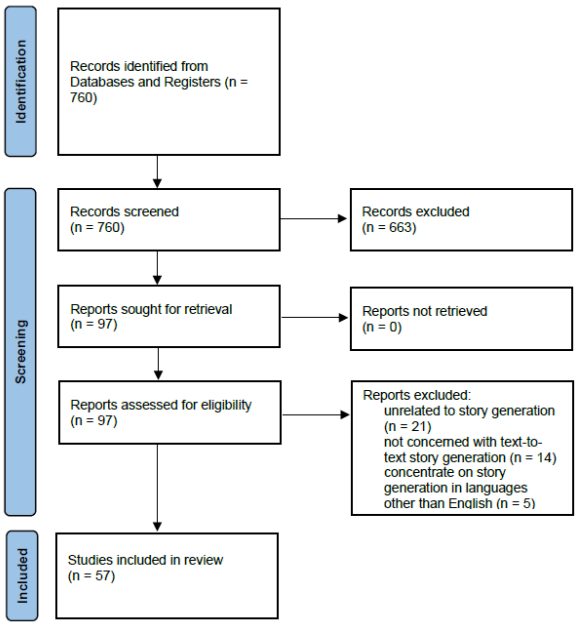

Figure 1: PRISMA chart results

Gómez-Rodríguez and Williams (2023) conduct an evaluation by instructing the models to produce epic-style narratives and then comparing these outputs with human-written counterparts. Human evaluations reveal that, although human authors surpass LLMs in terms of originality and humor, they fall behind the strongest models with respect to readability and adherence to epic-genre conventions.

In contrast, Marco et al. (2024) carry out a similar experiment comparing stories produced by a Generative Pre-trained Transformer (GPT; GPT-4 to be specific) and a professional novelist in an AI–human duel format. These works are assessed by literary experts, and the results reveal that GPT-4's stories were consistently rated lower across all quality dimensions. They conclude that LLMs still lack the nuanced depth, originality, and intentionality characteristic of top novelists.

Focusing on educational contexts, Valentini et al. (2023) examine the ability of several popular LLMs to generate stories with appropriate lexical and readability levels. The study finds that current models struggle to adjust their vocabulary to suit younger readers. They also evaluate the performance of state-of-the-art lexical simplification models in the children's story domain and they show that these models can simplify overly complex words after fine-tuning.

Later on, Marco et al. (2025) conduct a different experiment that examines fine-tuned Small Language Models (SLMs) like Bidirectional and Au-

Data Source	Search Target	Search Query
ACL Anthology	title, abstract	("natural language processing" OR "nlp" OR "language model" OR "llm") AND ("story generation" OR "storytelling") site:aclanthology.org
ACL, EMNLP, NAACL, TACL, COLING	title	("story" OR "stories" OR "fiction" OR "novel" OR "narrative" OR "writing")

Table 1: Search Strategy

toregressive Transformer (BART) against humans and LLMs. They show that SLMs can be competitive with both average humans and larger models in creative writing tasks, especially when flexibility is valued over strict consistency.

Recently, Tian et al. (2024) have compared humans and LLMs in story writing based on three discourse elements: story arcs at the macro level, turning points at the meso level, and affective dimensions at the micro level. They find that models lack narrative diversity and struggle to develop crucial turning points, such as major setbacks and climaxes, leading to less engaging stories.

4 Story Generation Methods

A significant portion of the reviewed articles centers on developing techniques to improve different aspects of story generation, such as coherence, consistency, interestingness, creativity, and controllability. This section provides an overview of the methodological approaches in automatic story generation by categorizing the studies into three technical groups based on model structure as follows: single-agent, multi-agent, and human-in-the-loop story generation.

4.1 Single-agent story generation

Single-agent story generation refers to approaches in which a single model or a unified generation pipeline is responsible for producing the entire narrative. Because most of the reviewed articles fall into this category, we further divide them into three groups based on the aspects they focus on: coherence, creativity, and controllability.

4.1.1 Coherence

Coherence is a foundational aspects of story generation, concerning how well the parts of a story fit together to form a consistent and logically sound narrative. A common approach to improve coherence is the use of hierarchical structures that begin with a high-level narrative structure, such as an outline or plan, and incrementally expand it into a complete story. This approach has been shown to substantially improve narrative coherence, especially in the generation of long-form texts (Fan et al., 2018).

One of the most influential studies in this area in recent years is Yang et al. (2022)'s LLM based framework. Their framework expands an initial premise into a detailed plan and iteratively drafts, revises, and edits stories using GPT-3. To overcome the limitations of their outlines often lacking specificity and not scaling effectively to longer texts, Yang et al. (2023) employs a hierarchical outlining process capable of recursively generating detailed outlines at multiple levels of granularity. Additionally, it incorporates a detailed controller to ensure faithfulness to the outline during story generation.

Besides Yang et al. (2022) and Yang et al. (2023), four other studies have explored hierarchical frameworks for story generation. Chen et al. (2022) introduce a contrastive soft prompt method that first trains a text representation aligned with coherent examples and distinguished from incoherent ones using contrastive learning. Then it applies this representation as a soft prompt to guide the generation process. Ma et al. (2023) propose a multi-stage model that leverages schema acquisition to incorporate structured knowledge into the plot generation process. Gandhi et al. (2023) present a scriptwriting workbench using GPT-3 fine-tuned on $1,000$ Hollywood movie scenes to generate both plots and scripts. Most recently, Li et al. (2025) have introduced a different LLM-based system, which uses subject–verb–object and subject–verb–subject triplets to construct plot structure, helping maintain logical flow and guide narrative generation. The system also incorporates a narrative entity knowledge graph to strengthen plot coherence.

Building upon these works, two studies have identified inherent limitations of LLMs that cannot be easily addressed through prompt engineering. Lei et al. (2024) observe that LLMs employed in prior frameworks often demonstrate insufficient

planning and linguistic capabilities for effective novel writing. To address these challenges, they propose to automatically construct storylines by learning from existing novels. Their framework first identifies structural information from existing novel datasets and then integrates this information to create a fine-tuning corpus for LLM adaptation to finally generate stories through a tree-like expansion process. Similarly, Wang et al. (2025a) note that LLMs inherently lack a deep understanding of storytelling principles and often suffer from memory limitations that lead to contextual inconsistencies. To mitigate these issues, they introduce a dynamic hierarchical outline mechanism, which guides story generation following the plan–write framework and established writing theories (Campbell, 1949). They also implement a memory enhancement module composed of temporal knowledge graphs to support planning and maintain narrative continuity.

In addition to hierarchical techniques, several studies propose alternative approaches improving reasoning and narrative logic. Peng et al. (2022b) achieve that by implementing a knowledge graph based reader model to reason about how the story should progress. Stories are generated based on the reader model, which is updated as new content is produced. Peng et al. (2022a) also investigate the role of commonsense inference in story generation, which leverages the Crosslingual Optimized Metric for Evaluation of Translation (COMET) model to infer a set of commonsense relations for each prompt sentence. These relations are then used to generate sentences and to verify whether commonsense relations remain consistent. Similarly aiming at improving coherence but using a different strategy, Brei et al. (2024) draw inspiration from the idea that strong narrative closure arises from well-aligned endpoints. Their framework generates the opening and closing sentences first, and then constructs the middle portion to ensure a coherent progression between the two. Likewise, Zhang and Long (2025) target coherence from another angle by incorporating actions and emotions as a mechanism to connect with a narrative arc (Boyd et al., 2020). Their method uses LLMs to detect logical holes in the narrative. Then, it combines character behaviors and emotions in the surrounding context to predict the possible actions and emotions and complete the missing plot.

Other studies seek to enhance coherence through capturing contextual features. To enable story generation given an initial context and event plan, Tang et al. (2022) use cross-attention to residually map contextual features to event sequences. Lu et al. (2023) also focus on events in their context. Their model uses a narrative-order–aware framework that employs a bidirectional pretrained model to encode event relationships and ordering with a BART-based generator that is fine-tuned using reinforcement learning. Pei et al. (2024) approach capturing context for enhanced coherence from a different angle. In their system, one LLM is responsible for generating the story content while another LLM, called an action discriminator, evaluates the current narrative state and selects the most suitable next action, guiding the progression of the story in a way where the actions should be context-sensitive and plot-progression coherent.

Researchers also study tasks that benefit downstream story generation. Paul and Frank (2021) focus on narrative story completion, using a fine-tuned GPT-2 model to infer contextual reasoning rules and iteratively generate the next sentence in a story. In contrast, Ma et al. (2024) investigate story premise synthesis, arguing that high-quality premises lead to stronger narratives. Their method decomposes a premise into multiple hierarchical modules and constructs a nested dictionary of consistent candidate elements. An LLM then selects and expands a key path within this structure to form a coherent premise, providing a semantically diverse foundation for story generation.

4.1.2 Creativity

Creativity concerns the richness, originality, and expressiveness of generated stories. Research in this area focuses on producing stories that are engaging, interesting, and capable of evoking emotional or imaginative responses from readers.

Huang et al. (2023) employ a GPT-based language model to generate base stories focusing on coherence, then applies Dynamic Beam Sizing and Affective Reranking to insert intriguing twists into the narratives to generate more empathy-evoking and interesting stories. Moreover, Park et al. (2025) boost creativity by generating visual representations of core story elements (e.g., characters, staging, and plot progression (Boyd et al., 2020)) and then producing multiple persona candidates based on these visuals, selecting the most suitable one to integrate into the narrative. Meanwhile, Wen et al. (2023) concentrate on increasing narrative complexity. Their model first creates a retrieval

repository containing multiple human-written stories and retrieves the most similar story to assist in generating the initial story. It then employs an "asking-why" prompting scheme that iteratively builds an evidence forest addressing ambiguities in the story.

4.1.3 Controllability

Controllability addresses the ability to guide the generation process toward specific attributes, constraints, or user-defined conditions. Instead of focusing solely on overall narrative quality, these studies aim to ensure that generated stories conform to predefined content requirements, such as themes, plots, characters, or stylistic features.

The first group of approaches incorporates explicit content constraints into the narrative. Wang et al. (2022) explore the creation of customized stories with characters, corresponding actions, and emotions arbitrarily assigned. Their model generates stories conditioned on previous content and a sequence of k fine-grained control conditions for the next sentence using BART. In a more targeted direction, Vijjini et al. (2022) aim to control interpersonal relationships within narratives. They employ a Bidirectional Encoder Representations from Transformers (BERT)-based relationship selector to determine which relationship should appear in the next sentence, followed by a GPT-2 story generator that continues the narrative accordingly. Rather than focusing on character control, Islam et al. (2024) explore generating stories that convey data characteristics. They introduce a multi-data story generation model with a planning module that extracts insights to form an outline and a narration module that produces and iteratively refines the full story using an LLM-based critic.

The second group of studies focuses on controlling higher-level, conceptual factors, such as psychological attributes. Kong et al. (2021) propose a model that generates stories in a specified style. Their system predicts a keyword distribution from the opening sentence and the desired style, and then uses those keywords to guide the rest of the narrative. Mori et al. (2022) propose combining a sequential language model and PPLM (Dathathri et al., 2019) to control emotion in story completion tasks. Xie et al. (2022), on the other hand, try to generate controllable stories that align with the story context and the protagonist's psychological state chains. They implement a state planner and trackers to memorize local psychological states and adapt them to obtain the protagonist's global psychological states for planning storytelling. Finally, a psychology controller integrates both local and global psychological states into the story context representation to compose psychology-guided stories using a BART-based model. Similarly, Zhang et al. (2022) focus on controlling the protagonist's persona in story generation. They achieve that by producing persona-related events and a sequence of keywords to guide story generation. They also use a dynamically expanding local knowledge graph to support plot generation.

4.2 Multi-agent Story Generation

Multi-agent story generation refers to approaches in which multiple models or agents work together to simulate story scenarios or jointly construct narratives, often leading to more dynamic and contextually rich storytelling.

Bae and Kim (2024) aim at boosting story creativity while preserving narrative coherence through collaborative LLM-based critics. Their method relies on a team of LLM-critics working with a leader model to iteratively refine both the story plan and the drafted narrative.

In contrast, Yu et al. (2025) and Ran et al. (2025) focus more in simulation. Yu et al. (2025) design a multi-agent story generation system in which a director agent coordinates character agents through roleplay, prompting them to act in ways consistent with the story outline, and then employs an LLM to transform these interactions into a story. In comparison, the simulation framework by Ran et al. (2025) first gathers character profiles and worldview information from source materials, and then employs LLM agents to enact scenes in which characters pursue their individual goals, ultimately generating a coherent story.

4.3 Human-In-The-Loop Story Generation

Human-in-the-loop story generation refers to approaches in which human authors and language models collaborate to produce narratives. Rather than relying solely on automated generation, these studies investigate how AI can assist, guide, or augment human creativity throughout the storytelling process, supporting co-creative writing experiences.

One group of studies focuses on improving story quality. Rosa et al. (2022) develop a GPT-2–based interactive system for human-in-the-loop theatre script generation, using a two-phase hierarchical

method to enhance output quality. The system generates the script line by line, allowing users to intervene by regenerating previous lines or choosing the next speaker. Meanwhile, Zhong et al. (2023) enhance story quality by incorporating writing modes as a control mechanism. Their model employs a fine-tuned language model to generate text in specific writing styles (Dialogue, Action, and Description) with a classifier selecting the appropriate mode to extend the narrative. On the other hand, Ermolaeva et al. (2024) propose a nonlinear fairy-tale generation pipeline where users define a protagonist and select or modify suggested actions. Their system uses prompt engineering to enforce emotion across the narrative arc that alternates between "low" and "rise" phases, balancing setbacks and positive developments until the protagonist's goal is reached.

The second group of studies focuses on user experience. Lee et al. (2022) introduce an interface design that supports children and parents in collaboratively rewriting stories using a GPT-2-based system. The model first identifies entities in the story that can be modified based on a set of predefined dimensions and then generates questions for parents to ask their children. The story is subsequently rewritten according to the children's answers, encouraging creativity and interactive learning. Similarly, Saraswat et al. (2024) present an interactive story creation platform for children. The system constructs a customized knowledge graph from a dataset of children's stories and integrates it with an LLM to collaboratively generate new narratives, combining structured knowledge with generative creativity. Lee and Chang (2023) extend to English language learning at schools by designing a dialogue-based story co-telling module aimed at enhancing English narrative skills among English as a Second Language (ESL) learners. The system utilizes knowledge graphs to comprehend the storyline, while two agents generate dialogue responses informed by dialogue history and the knowledge graph and trained using reinforcement learning. ESL-learners interact by selecting which agent's response to include in the story. In contrast, Arnold (2023) implements a gamified, quiz-based classroom approach across two university courses on NLP. In this framework, questions related to lecture content are presented through story-driven narratives, allowing students to answer in a dynamic and competitive setting.

Dataset	No.
ROCStories (Mostafazadeh et al., 2016)	13
WritingPrompts (Fan et al., 2018)	12
CMU Movie Summary Corpus (Bamman et al., 2013)	2
Story Commonsense (Rashkin et al., 2018)	2
Fairy tales (Ammanabrolu et al., 2020)	2

Table 2: Number of Papers Using Each Dataset in Automatic Story Generation

5 Dataset

We find five datasets that have each been used in more than one of the surveyed publications (Table 2). Two of these, ROCStories (Mostafazadeh et al., 2016) and WritingPrompts (Fan et al., 2018) are by far the most frequently used. In addition, several new datasets have recently been introduced.

- **ROCStories** (Mostafazadeh et al., 2016) — The ROCStories dataset contains $98,161$ human-written English stories, each composed of five sentences.
- **Writing Prompts** (Fan et al., 2018) — A large-scale dataset of approximately $300,000$ human-written stories paired with writing prompts sourced from Reddit. The average story length is 59.35 sentences.
- **Fairy tales** (Ammanabrolu et al., 2020) — A collection of 695 fairy-tale-style stories extracted from Wikipedia story summaries, with an average length of 24.8 sentences per story.
- **Story Commonsense** (Rashkin et al., 2018) — A dataset of $4,853$ five-sentence stories annotated with characters' emotions and motivations.
- **CMU Movie Summary Corpus** (Bamman et al., 2013) — A large corpus of over $42,000$ movie plot summaries and related metadata, compiled by researchers at Carnegie Mellon University (CMU).

Tikhonov et al. (2021) introduce a multilingual dataset where stories and characters are cross-linked across languages and annotated by genre and topic, with data scraped from Wikipedia. To explore the role of narrative in argumentation, Falk and Lapesa (2023) develop a dataset derived from corpora in computational argumentation and

Evaluation Metric	No.
Human evaluation	32
BLEU-n (Papineni et al., 2002)	17
ROUGE-N/ROUGE-L (Lin, 2004)	10
LLM-based Evaluation	12
Perplexity	8
Distinct-n (Li et al., 2016)	8
BERTScore (Zhang et al., 2019)	8
Repetition-n (Shao et al., 2019)	5
UNION (Guan and Huang, 2020)	4

Table 3: Number of Papers Using Each Evaluation Metric in Automatic Story Generation

the social sciences. It includes annotated textual spans labeled for argumentative functions and narrative properties. To investigate collaborative storytelling with LLMs, Du and Chilton (2023) introduce a collection of collaboratively written stories from `storywars.net`. The dataset also includes a benchmark with seven understanding and five generation tasks. Expanding story generation to domain-specific knowledge, Jiang et al. (2024) develop a dataset for legal education that includes legal concepts, their definitions, LLM-generated stories, questions, and human annotations. To address moral dimensions in storytelling, Guan et al. (2022) create a dataset consisting of human-written stories in Chinese and English paired with moral annotations. To evaluate LLMs' ability to generate both generic and personalized narratives based on predefined morals and identity elements, Yunusov et al. (2024) propose a corpus of personalized short stories that incorporate user identity traits.

6 Evaluation methods

We find nine evaluation metrics that have each been used in more than one of the surveyed publications (Table 3). Human evaluation remains the most widely used metric in story generation research. However, there is a growing trend toward using LLMs as substitutes for human evaluators.

- **Human Evaluation** — Human judges manually assess or compare generated stories, often by assigning scores, conducting pairwise comparisons, or providing comments.
- **BLEU-n** (Papineni et al., 2002) — Bilingual Evaluation Understudy (BLEU) measures story quality based on the degree of n-gram overlap between a generated story and a human-written reference.
- **ROUGE-N/ROUGE-L** (Lin, 2004) — Recall-Oriented Understudy for Gisting Evaluation (ROUGE) refers to a set of measures. ROUGE-N measures the number of matching n-grams between the model-generated text and a human-produced reference, while ROUGE-L evaluates the longest common subsequence.
- **LLM-based Evaluation** — A LLM is used to assess generated stories in a human-like manner, offering automated qualitative judgment.
- **Perplexity** — Perplexity measures the uncertainty of generated tokens predicted by neural models.
- **Distinct-n** (Li et al., 2016) — Distinct-n computes the ratio of unique n-grams to all generated n-grams, measuring text diversity.
- **BERTScore** (Zhang et al., 2019) — BERTScore compares generated and reference texts by aligning their contextual embeddings using cosine similarity.
- **Repetition-n** (Shao et al., 2019) — Repetition-n measures redundancy by calculating the proportion of generated stories that contain at least one repeated n-gram.
- **UNION** (Guan and Huang, 2020) — A learnable metric that employs a classifier trained on human-written and perturbed texts.

Beyond commonly used metrics, several new benchmarks and evaluation methods have been proposed to assess story generation more effectively. Clark and Smith (2021) present a collaborative framework for pairwise model evaluation. In this setup, two models provide alternative suggestions to participants as they write short stories. After completing the story, writers provide feedback on their experience and the quality of the model-generated suggestions. In contrast, Chhun et al. (2022) enhance the human evaluation framework by designing a set of non-redundant criteria for assessing automatic story generation. They introduce a large human-annotated benchmark comprising stories from the WritingPrompts (Fan et al., 2018) dataset, each rated by three annotators.

Subsequent research has focused on developing frameworks that leverage language models for automatic evaluation. Yang and Jin (2025) introduce a model for efficient summary-based reviewing that evaluates stories through plot, character, and writing analyses. They also build a benchmark of books with ratings and reader reviews. Chen et al. (2023) present a human preference-liked evaluation framework with three subtasks: Ranking, Rating, and Reasoning. To aid their system, they construct a new story dataset by crowd-sourcing paired ranked stories from a writing prompt website and annotated by crowd workers on Amazon Mechanical Turk. Wang et al. (2025b) propose an annotated fiction dataset in English and Chinese, and design a multi-level evaluation framework using LLM based on ten metrics, such as creativity and grammaticality, across macro, meso, and micro levels.

Other efforts involve using negative samples to aid in evaluating story coherence and quality. Guan et al. (2021) introduce a benchmark for assessing open-ended story generation metrics. It includes a manually annotated dataset and an automatically constructed dataset producing negative samples designed to test metric robustness and coherence evaluation. Ghazarian et al. (2021) develop an approach for generating more realistic negative samples by introducing plot-level incoherence to guide models in producing implausible yet challenging examples, filtered using adversarial techniques. Xie et al. (2023) measure story quality by comparing the likelihood difference between original and perturbed versions, based on the idea that higher quality stories will exhibit more significant effects from the perturbation compared to lower quality ones.

7 Challenges

Based on the examined literature, we identified two main persistent challenges in story generation.

One of it lies in evaluation. Although various benchmarks and metrics have been proposed, the field still lacks a standardized, universally accepted evaluation framework. This gap makes it difficult to conduct experiments systematically or compare models reliably. The problem is further compounded by the limitations of automatic evaluation methods. While surface-level attributes such as length and diversity can be measured easily, conceptual qualities like coherence and interestingness remain difficult to assess. Although recent studies have explored using LLMs to approximate human judgments on these aspects, issues such as poor reproducibility persist.

Another major challenge in story generation is maintaining coherence in long-form narratives. While LLMs have substantially improved local coherence, preserving global consistency remains difficult, primarily due to the inherent memory limitations of current models. This challenge is further compounded by the scarcity of suitable training datasets. The two most widely used corpora, WritingPrompts (Fan et al., 2018) and ROCStories (Mostafazadeh et al., 2016) are too short to support the development of models aimed at producing long-form stories. Although new datasets would help bridge this gap, their creation is constrained by copyright restrictions, as authors are often unwilling to release their work for training purposes, and publicly available stories tend to be outdated or low quality.

8 Discussion

In this systematic review, we presented a comprehensive examination of story-generation research published over the past four years. Through analyzing the papers identified in our search, we observed a clear shift in methodology driven by the adoption of LLMs, alongside persistent challenges, such as evaluation and maintaining coherence in long-form narratives, that continue to limit progress.

Compared with surveys published before 2024 (Alhussain and Azmi, 2021; Fang et al., 2023), analogously to Teleki et al. (2025), we observe a distinct shift toward the use of LLMs. In previous work, structural models and planning-based approaches are still considered two major branches of story generation. In contrast, although our review does not restrict itself only to language model based methods, every study identified through our search incorporates Transformer-based language models in some capacity, with more than half relying specifically on LLMs. Notably, we identify no studies included in our review that rely exclusively on structural or planning-based frameworks.

On the other hand, despite significant progress in modeling, advances in developing datasets and evaluation metrics remain limited. ROCStories (Mostafazadeh et al., 2016) and WritingPrompts (Fan et al., 2018) continue to be the most commonly used datasets, and although several new datasets have been introduced, none has yet achieved broad adoption within the research community. Accord-

ing to our review findings, evaluation metrics face similar challenges. Human evaluation remains the most reliable and widely accepted approach. A growing number of studies have begun to use LLMs as substitutes for human judges, yet there is still no unified rubric or standardized procedure for conducting LLM-based evaluation effectively.

Based on our findings, we propose the following three future directions for advancing the field:

1. Developing new evaluation metrics. Reliable automatic evaluation methods would not only support model comparison and effective model training but also more meaningful engagement of humans in evaluation studies; a conclusion made by Hämäläinen and Alnajjar (2021) too. Although human evaluation remains indispensable in the absence of dependable automatic metrics, expert evaluators are rarely given comprehensive rubrics and commenting options for each assessment dimension, or their valuable inputs are not fully utilized in evaluation studies. Future research should adopt and experiment with newly proposed metrics consistently in order to support method comparisons, track performance enhancements in time, and meaningfully engage with human experts.

2. Creating more suitable datasets. Among the five commonly used datasets we identified in Table 2, the longest story is up to a few thousand words, still insufficient for models designed to generate novel-length narratives. Furthermore, all of these datasets are sourced from nonprofessional writers, resulting in arguably inconsistent quality. Finally, their data contamination with LLMs is expected to cause evaluation issues (Chen et al., 2025; Xu et al., 2025).

3. Ongoing improvement of narrative coherence and consistency and their exploration in practical applications of story generation. Developing methods that allow models to efficiently retain, retrieve, and reason over previously generated content will still be the key for producing long-form, high-quality narratives. As current models demonstrate strong performance only on short stories, it is also important to investigate how these techniques can be applied in real-world contexts (Section 4.3) and extended to longer coherent and consistent narratives (Section 7). This, in turn, will allow studying how these applications and extensions can inform putting automatic long-story generation into practice. Grounding story-generation tasks and evaluations in realistic scenarios would increase the practical relevance of the tasks and clarify how these techniques can provide real-world benefits.

In conclusion, by analyzing 57 studies in text-to-text story generation, we demonstrate that the widespread adoption of Transformer architectures and LLMs has substantially improved narrative quality in automatic story generation. Nevertheless, challenges remain, particularly in developing robust evaluation metrics and maintaining coherence in long-form narratives. To support future research, we propose developing new evaluation metrics and creating more suitable datasets, together with ongoing improvement of narrative coherence and consistency, as well as their exploration in practical applications. We hope that our synthesis provides a comprehensive foundation for guiding the next generation of studies in story generation.

Limitations

This study has several limitations. First, our coverage is restricted to venues within computational linguistics and NLP. In addition, we exclude research on text-to-image or image-to-text story generation, as well as work focused on languages other than English. As a result, some evaluation strategies and methodological approaches may not be captured. Second, due to page limits, we are unable to provide detailed explanations of every method identified, and we focus on summarizing the datasets and techniques that appear most frequently in the paper. There also exist datasets and evaluation metrics that are used only once or in a single paper, which we do not discuss in depth.

Acknowledgment

We thank The Australian National University (ANU) and the ANU School of Computing for supporting the PhD studies of the first two authors. We also express our gratitude to the anonymous reviewers for their helpful comments.

References

Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Comput. Surv.*, 54(5).

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 3–9.

Nantheera Anantrasirichai and David Bull. 2022. Artificial intelligence in the creative industries: a review. *Artificial intelligence review*, 55(1):589–656.

Thomas Arnold. 2023. Quest: Quizzes utilizing engaging storytelling. In *Proceedings of the 1st Workshop on Teaching for NLP*, pages 28–36.

Minwook Bae and Hyounghun Kim. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

Ryan L. Boyd, Kate G. Blackburn, and James W. Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

Anneliese Brei, Chao Zhao, and Snigdha Chaturvedi. 2024. Returning to the start: Generating narratives with related endpoints. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 101–112.

Joseph Campbell. 1949. *The hero with a thousand faces*.

Guandan Chen, Jiashu Pu, Yadong Xi, and Rongsheng Zhang. 2022. Coherent long text generation by contrastive soft prompt. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 445–455.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2023. Storyer: Automatic story evaluation via ranking, rating and reasoning. *Journal of Natural Language Processing*, 30(1):243–249.

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10091–10109, Suzhou, China. Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836.

Elizabeth Clark and Noah A Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Yulun Du and Lydia Chilton. 2023. StoryWars: A dataset and instruction tuning baselines for collaborative story understanding and generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3044–3062, Toronto, Canada. Association for Computational Linguistics.

Marina Ermolaeva, Anastasia Shakhmatova, Alina Nepomnyashchikh, and Alena Fenogenova. 2024. How to tame your plotline: A framework for goal-driven interactive fairy tale generation. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 8–31.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Xiaoxuan Fang, Davy Tsz Kit Ng, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2023. A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies*, 28(11):14361–14397.

Prerak Gandhi, Vishal Pramanik, and Pushpak Bhattacharyya. 2023. Kurosawa: A script writer's assistant. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 540–550, Goa University, Goa, India. NLP Association of India (NLPAI).

Sarik Ghazarian, Zixi Liu, Ralph Weischedel, Aram Galstyan, Nanyun Peng, and 1 others. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528.

Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95, Online. Association for Computational Linguistics.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11792–11806.

Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Datanarrative: Automated data-driven storytelling with visualizations and texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19253–19286.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and 1 others. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436.

Yoonjoo Lee, Tae Soo Kim, Minsuk Chang, and Juho Kim. 2022. Interactive children's story rewriting through parent-children interaction. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 62–71.

Yu-Kai Lee and Chia-Hui Chang. 2023. Story co-telling dialogue generation based on multi-agent reinforcement learning and story highlights. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 11–19.

Huang Lei, Jiaming Guo, Guanhua He, Xishan Zhang, Rui Zhang, Shaohui Peng, Shaoli Liu, and Tianshi Chen. 2024. Ex3: Automatic novel writing by extracting, excelsior and expanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9125–9146.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 598–604.

Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, and 1 others. 2025. Storyteller: An enhanced plot-planning framework for coherent and cohesive story generation. *arXiv preprint arXiv:2506.02347*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and 1 others. 2023. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Congda Ma, Kotaro Funakoshi, Kiyoaki Shirai, and Manabu Okumura. 2023. Coherent story generation with structured knowledge. In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 681–690.

Yan Ma, Yu Qiao, and Pengfei Liu. 2024. Mops: Modular story premise synthesis for open-ended automatic story generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2135–2169.

Guillermo Marco, Julio Gonzalo, M Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024. Pron vs prompt: can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19654–19670.

Guillermo Marco, Luz Rello, and Julio Gonzalo. 2025. Small language models can outperform humans in short creative writing: A study comparing slms with humans and llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6552–6570.

Natalie Merrill and Robyn Fivush. 2016. Intergenerational narratives and identity across development. *Developmental Review*, 40:72–92.

Yusuke Mori, Hiroaki Yamane, Ryohei Shimizu, and Tatsuya Harada. 2022. Plug-and-play controller for story completion: A pilot study toward emotion-aware story writing assistance. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 46–57.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kyeongman Park, Minbeom Kim, and Kyomin Jung. 2025. A character-centric creative story generation via imagination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1598–1645.

Debjit Paul and Anette Frank. 2021. Coins: Dynamically generating contextualized inference rules for narrative story completion. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5086–5099.

Jonathan Pei, Zeeshan Patel, Karim El-Refai, and Tianle Li. 2024. Swag: Storytelling with action guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14086–14106.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022a. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029.

Xiangyu Peng, Kaige Xie, Amal Alabdulkarim, Harshith Kayam, Samihan Dani, and Mark Riedl. 2022b. Guiding neural story generation with reader models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7087–7111.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299.

Rudolf Rosa, Patrícia Schmidtová, Ondřej Dušek, Tomáš Musil, David Mareček, Saad Obaid, Marie Nováková, Klára Vosecká, and Josef Doležal. 2022. Gpt-2-based human-in-the-loop theatre play script generation. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 29–37.

Hryadyansh Saraswat, Snehal D Shete, Vikas Dangi, Kushagra Agrawal, Anuj Aggarwal, and Aditya Nigam. 2024. Story-yarn: An interactive story building application. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 248–255.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3257–3268.

Wendy A Suzuki, Mónica I Feliú-Mójer, Uri Hasson, Rachel Yehuda, and Jean Mary Zarate. 2018. Dialogues: The science and power of storytelling. *Journal of Neuroscience*, 38(44):9468–9470.

Chen Tang, Chenghua Lin, Henglin Huang, Frank Guerin, and Zhihao Zhang. 2022. Etrica: Event-triggered context-aware story generation augmented by cross attention. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5504–5518.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and 1 others. 2025. A survey on llms for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*,

pages 32–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598.

Anvesh Rao Vijjini, Faeze Brahman, and Snigdha Chaturvedi. 2022. Towards inter-character relationship-driven story generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8970–8987.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025a. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391.

Wenqing Wang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025b. Towards a "novel" benchmark: Evaluating literary fiction with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21648–21673.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. Chae: Fine-grained controllable story generation with characters, actions and emotions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6426–6435.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. Grove: A retrieval-augmented complex story generation framework with a forest of evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998.

Yuqiang Xie, Yue Hu, Yunpeng Li, Guanqun Bi, Luxi Xing, and Wei Peng. 2022. Psychology-guided controllable story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6480–6492.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Lau. 2023. Deltascore: Fine-grained story evaluation with perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5317–5331.

Cheng Xu, Nan Yan, Shuhao Guan, Changhong Jin, Yuke Mei, Yibing Guo, and Tahar Kechadi. 2025. DCR: Quantifying data contamination in LLMs evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23013–23031, Suzhou, China. Association for Computational Linguistics.

Dingyi Yang and Qin Jin. 2025. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16375–16398.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479.

Tian Yu, Ken Shi, Zixin Zhao, and Gerald Penn. 2025. Multi-agent based character simulation for story writing. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 87–108.

Sarfaroz Yunusov, Hamza Sidat, and Ali Emami. 2024. Mirrorstories: Reflecting diversity through personalized narrative generation with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6702–6717.

Jinming Zhang and Yunfei Long. 2025. Mld-ea: Check and complete narrative coherence by introducing emotions and actions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1892–1907.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv e-prints*, pages arXiv–2403.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist's persona in story generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3346–3361.

Wenjie Zhong, Jason Naradowsky, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023. Fiction-writing mode: An effective control for human-machine collaborative writing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1752–1765.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and

Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.