A Survey on LLMs for Story Generation

Maria Teleki, Vedangi Bengali*, Xiangjue Dong*, Sai Tejas Janjur*, Haoran Liu*, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, James Caverlee
Texas A&M University,
Correspondence: mariateleki@tamu.edu

Abstract

Methods for story generation with Large Language Models (LLMs) have come into the spotlight recently. We create a novel taxonomy of LLMs for story generation consisting of two major paradigms: (i) independent story generation by an LLM, and (ii) author-assistance for story generation – a collaborative approach with LLMs supporting human authors. We compare existing works based on their methodology, datasets, generated story types, evaluation methods, and LLM usage. With a comprehensive survey, we identify potential directions for future work.

1 Introduction

Highly capable LLMs like ChatGPT, Llama, and more (Achiam et al., 2023; Grattafiori et al., 2024) open up possibilities to rethink and re-formulate the static, existing ways of storytelling (Choo et al., 2020). For example, **with LLMs, stories can be interactive (Wang et al., 2024) and personalized (Lee et al., 2024)**, responding flexibly to users in real time. These new ways of storytelling create significant economic opportunities, for example: improving player experiences in the gaming industry (Wang et al., 2024), improving childcare quality and training health professionals in the healthcare industry (Moreau et al., 2018), improving teaching methods in education (Robin, 2008; Ohler, 2006), and improving movie script development in the entertainment industry (Dayo et al., 2023).

Despite the appealing capabilities of LLMs, LLM outputs often suffer from hallucinations, factual inaccuracies, and the generation of offensive content. Furthermore, existing models may be incapable of conceptualizing key story arcs and understanding and interpreting nuanced human emotions. For example, Subbiah et al. (2024) finds that LLMs are not able to correctly summarize key aspects of stories, including the story subtext and the (un)reliability of story narrators.

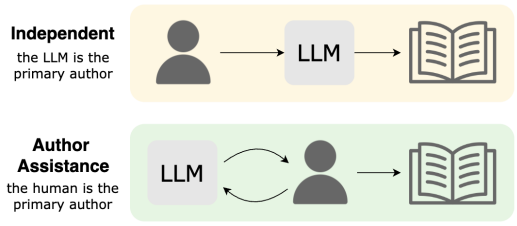

Figure 1: **Major Taxonomy Categories**: In our taxonomy (Figure 2), we first categorize works based on primary authorship.

While there is a vast literature of work on digital storytelling (Trichopoulos et al., 2023; Wu and Chen, 2020; Ohler, 2006) and the use of traditional language models – i.e., pre-LLM[1] – for story generation (Alhussain and Azmi, 2021; Fang et al., 2023), there is a gap in the literature in surveying the use of LLMs in story generation. The closest prior work, Li et al. (2024), surveys storytelling for data interpretation applications, whereas we focus our work on storytelling in the traditional sense (i.e., non-data-centric storytelling), spanning application areas from education to the interactive video game story generation. Specifically, we survey recent early-stage works using LLMs[1] for story generation to close the gap.

We provide a systematic understanding of the area to highlight the opportunities for follow-up work. We bridge the gap between HCI-style story systems and NLP-style story systems, ideating future work including: the creation of large-scale datasets and metrics, the use of open-sourced and small models, and the use of inference-time methods for effectively controlling LLMs (Welleck et al., 2024; Dong et al., 2024). We make the following contributions:

- We introduce a novel **Taxonomy of LLMs for Story Generation** (Figure 2, §2), categorizing

[1]Here, we consider LLMs as Large Language Models released after 2022, following the emergence of GPT-4.

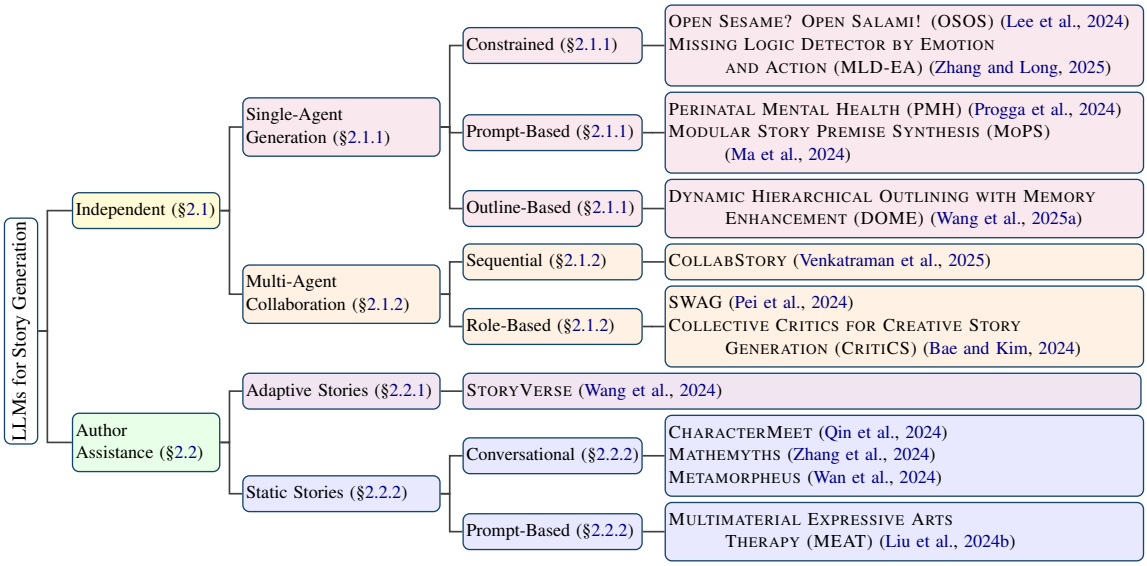

Figure 2: Taxonomy of LLMs for Story Generation. *Note that, for clarity, unnamed frameworks are assigned descriptive labels reflecting their key contributions.*

recent methods from top-tier venues.

- We conduct **a comprehensive comparison of these methods** in terms of datasets (§3), evaluation (§5), and LLM use (§3).
- We suggest **directions for future work** (§6).
- We release an **online community resource**: https://github.com/mariateleki/Awesome-Story-Generation.

2 Taxonomy

Our first contribution is a novel taxonomy of LLMs for Story Generation, shown in Figure 2. Our taxonomy divides story generation into two paradigms based on primary authorship: Independent (§2.1) and Author Assistance (§2.2). Independent story generation methods consider the LLM to be the primary author. This is in contrast with Author Assistance methods, which consider the primary author to be a human author, and the LLM acts as an assistant in an interactive paradigm. Within these main categories, we further subdivide the work based on the most defining feature of their approach. The criteria and venues for paper selection in this survey are provided in Appendix A.

2.1 Independent

Independent story generation methods position the LLM as the primary author. We divide the approaches for independent story generation into methods that constrain story generation and more open-ended prompt-based methods.

2.1.1 Single-Agent Generation

Constrained. Constrained generation methods for LLM-based story generation encourage certain criteria to be met in their generations. These constraints may be driven by pedagogical goals, logical coherence, or consistency in narrative elements.

Lee et al. (2024) propose OPEN SESAME? OPEN SALAMI! (OSOS), a method for generating stories to help children practice vocabulary words which they struggle with. OSOS has three modules: (i) the Profiler, (ii) the Extractor, and (iii) the Generator. The Profiler takes audio input from the child's home and converts the audio to diarized text. The Extractor, then, is responsible for selecting the prioritized words, W_{all}, which it does via a linear combination of three important features: frequency, commonality across time, location, and speaker, and perceptual saliency, a measure of speech clarity. The top k words are selected to form $W_{>k}$, the set of the most prioritized words. Finally, the Generator is used to construct the story based on an existing abstract with $W_{>k}$. The generation process has multiple steps: (1) an initial story which incorporates $W_{>k}$ is generated based off an existing abstract using GPT-4, (2) a human reviews this story, (3) GPT-4 is used to pagnate the story, (4) Stable Diffusion is used to generate an image for each page of the story. A human-in-the-loop approach is utilized to make three checks throughout this process: (1) a web-based UI allows the user to steer the direction of the generated story, with prompts like "add more characters", "add more dialogue", and "add more conflicts", (2) a check on the image generations, and (3) a check on the final story. These checks position the human as an assistant to the LLM. One important issue noted with this system is character portrayal inconsistency across the

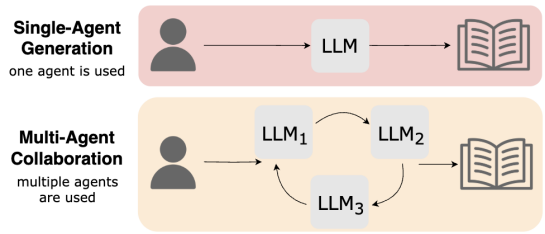

Figure 3: **Independent Generation Categories** (§2.1)

image generations – i.e., the main character looks different in different images.

Zhang and Long (2025) propose MISSING LOGIC DETECTOR BY EMOTION AND ACTION (MLD-EA), a method to improve the logical and emotional flow of generated stories. For each character and sentence in a story, MLD-EA breaks the sentence into actions, which are each classified with an emotion. The emotional categories are based on a psychological framework, and a null emotion is included. MLD-EA then predicts the indices, k, at which there is a logical flaw. These flaws were synthetically created by removing sentences from the original stories in the dataset. The emotion-action sequences at $k-1$ and k are then used for zero-shot sentence generation to generate a sentence that logically bridges the formerly illogical sentences together. MLD-EA mainly relies on handcrafted templates for each module – i.e., identifying k and then generating new sentences to logically string the story together. The authors evaluate their EA module on the *missing sentence prediction task*, and find that the EA module is helpful for predicting the next sentence. This system enhances the logical and emotional coherence of generated stories, therefore addressing a critical challenge in automated storytelling.

Comparing the methods, each method is designed to satisfy its constraints in different ways. OSOS generates the story, and focuses on constraint satisfaction (the inclusion of learning-targeted vocabulary words) via prompt-based methods with a human-in-the-loop approach for verification. MLD-EA focuses on correcting an existing story, finding logical gaps, and correcting them. The incorporation of emotion-action modeling represents a significant step toward more human-like narrative generation, where characters' decisions and story outcomes are influenced by plausible emotional and behavioral dynamics, and the story itself has a continuous and understandable plot.

Prompt-Based. Prompt-based methods utilize a zero-shot prompt to create stories and have no refinement or multi-module steps to further improve the generated story. These prompts typically focus on crafting broad narrative instructions without explicit modular feedback, differing from static story-assistance prompts where human-authored context and iterative refinement shape the prompt structure. These approaches emphasize simplicity and direct generation.

Progga et al. (2024) propose PERINATAL MENTAL HEALTH (PMH), a method to generate stories about perinatal mental health struggles, for the purpose of supporting maternal health via emotional resonance. A dataset of first-person experiences is analyzed via topic modeling, and then these topics are included to prompt the LLM to generate new experiential narratives of perinatal mental health (N=45 new stories). A qualitative analysis of the stories reveals that these stories largely adhere to the prompt specifications (38/45). However, there were some concerning recurring issues in the stories: detailed analysis reveals that there were hallucination issues, and that certain topics (e.g. *rape*, *harassment*) were sometimes met with refusal by the LLM.

Ma et al. (2024) introduce Modular Story Premise Synthesis (MoPS), a method for automatic story premise generation. MoPS breaks a premise into sequentially dependent modules, including theme, background, persona, and plot. LLMs generate candidate elements for each module, using outputs from previous modules as preconditions. Then, a key path is sampled, and the LLM synthesizes the selected elements into a compact, coherent premise. Both human evaluation and automated metrics are used to assess the diversity and quality of the generated premises. Results indicate that high-quality MoPS premises can effectively guide long story generation by incorporating a richer set of components, such as backgrounds and personas.

Outline-Based. Wang et al. (2025a) proposes DOME, Dynamic Hierarchical Outlining with Memory-Enhancement long-form story generation method, which combines structured planning with dynamic memory mechanisms. Central to this approach is the Dynamic Hierarchical Outline (DHO), which integrates narrative theory into the outline generation process and closely couples planning with writing. This fusion helps maintain plot coherence and completeness while allowing flexibility

to address uncertainties during generation. Additionally, a Memory-Enhancement Module (MEM), utilizing temporal knowledge graphs, captures and recalls previously generated content, thereby reducing contradictions and enhancing narrative consistency. To assess coherence, a Temporal Conflict Analyzer is employed, which automatically evaluates contextual alignment based on temporal relationships in the story.

2.1.2 Multi-Agent Collaboration

Multi-Agent Collaboration methods explore LLM-LLM collaboration in story generation. These agents can either contribute equally as co-authors or each LLM can perform a specific role in the writing process.

Sequential. In this framework, two or more LLMs work together as authors and iteratively build parts of the story. Each model takes turns sequentially adding the next segment, like plot twists, dialogues, or more scenic details, etc., based on the context generated so far. This helps enhance creativity in narratives, since no single agent is in complete control.

Venkatraman et al. (2025) proposes COLLABSTORY. This study focuses on long-form stories in various genres written by either single agents or up to 5 agents. Each agent writes a segment of the story and passes it to the next agent to add their own part, and the process continues until a coherent narrative is produced. By using different agent order permutations, they compile over 32,000 stories generated using open source and instruction-tuned LLM models. Evaluation studies show that multi-agent collaborations create more human-level stories as opposed to standalone agents. Additionally, this work also adapts the PAN authorship-analysis suite to a multi-agent setting and raises certain ethical concerns regarding the authorship credits, academic integrity, and use of malicious agents in spreading incorrect information. This system investigates the dynamics of multi-agent authorship, offering insights into how diverse LLMs can contribute distinct narrative styles and content.

Role-Based. In Role-Based multi-agent architectures, every AI agent performs a distinct function in the storytelling process. In contrast to the previous methods, not all agents take part in writing parts of the story. Instead, some agents can act as "content writers" while others can take roles like "high-level plot planners", or as "feedback models", etc. This division of responsibilities can help in storytelling with better control over the narrative style.

Pei et al. (2024) introduce SWAG: STORYTELLING WITH ACTION GUIDANCE, a flexible framework to generate long-form stories that uses a feedback loop to guide the narrative, framing storytelling as a search problem where the system iteratively selects the most contextually appropriate actions to advance the narrative. It consists of a story generation model (π_{story}) that writes the story content and an action-discriminator LLM model (π_{AD}) that selects the next best 'action' to drive the story's future direction. The process starts with a story prompt where π_{story} writes the first paragraph. The π_{AD} receives the current story state and a curated list of 30 high-level actions (for e.g. *add suspense*, *add plot twist*, etc.) from which it chooses the most engaging action and prompts back the π_{story} to write the next part of the story according to the suggested action. This iterative process continues to build the story step by step. The model is flexible in the sense that AD LLM can be used with any other LLMs for story generation, and various story genres can be targeted by customizing the list of actions. Various machine and human evaluations show the effectiveness of using the feedback model to generate more engaging and interesting stories without affecting their coherence. This approach signifies a shift toward more controlled and purposeful story generation, where LLM systems can self-regulate to produce more compelling narratives.

Bae and Kim (2024) propose CRITICS, a framework which generates stories via a pipeline of LLMs each prompted with a specific role to act as a critic. There are two major stages in the pipeline: (i) CRPLAN takes the user's input outline and uses a set of story-specific personas as critics in a multi-round process to produce a refined outline, assessing based on the following creativity criteria: *original theme and background setting*, *unusual story structure*, and *unusual ending*. An evaluator critic determines which edits to accept. (ii) CRTEXT which takes the story generated via the refined outline, and focuses on enhancing expressiveness-related aspects of the story – i.e., onomatopoeia and imagery. This approach represents a way to automate creativity-related story efforts.

2.2 Author Assistance

In contrast to independent story generation, where LLMs act as the primary author, author assistance

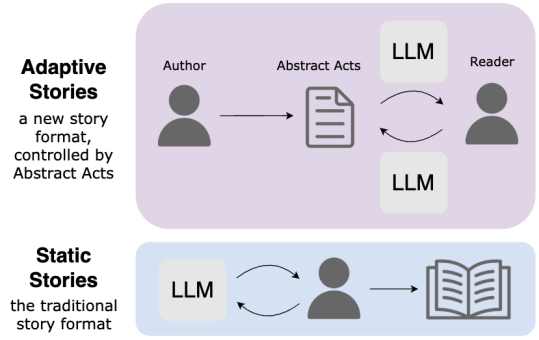

Figure 4: **Author Assistance Categories** (§2.2)

methods focus on supporting a primary human author in story creation. These methods use LLMs as part of their support tools and can be classified as adaptive or static in nature.

2.2.1 Adaptive Stories

Adaptive stories are not final-product stories once the author is finished writing – i.e., they do not take the form of a fixed string, but instead a variable string depending on reader inputs.

Wang et al. (2024) propose StoryVerse, a system that translates author-defined plot points – "**abstract acts**" – into detailed character actions, allowing for dynamic story evolution that still respects the author's plot plan. The method is used to create stories that are responsive to player actions in a video game. StoryVerse is comprised of two main modules: (i) an Act Director, and (ii) a Character Simulator. The Act Director intakes information about player actions in the video game environment, and the abstract acts from the author – e.g., *Character X goes on a vacation to Florida; Pre-Requisite: Character X falls in love with Character Y; Placeholder: None*. These abstract acts indicate *authorial intent*, imposing constraints on the generated character action sequences so that the story will play out in the way that the author intended. StoryVerse's approach exemplifies the potential of LLMs to augment human creativity, providing tools that enhance rather than replace the author's storytelling capabilities. This balance between control and emergence enables the creation of "living stories" that are still faithful to the authorial intent.

2.2.2 Static Stories

Static stories are final-product stories once the author is finished with them – i.e., they take the form of a fixed string. These stories do not change once they are composed.

Conversational. These stories are generated via a back-and-forth style conversation between the human author and the LLM-powered author assistance system. In these works, the LLM acts as part of a *creative support system*. The end product is still a fixed-string story. These systems offer a chat-style or brainstorming interface to assist human authors in the story creation process.

Qin et al. (2024) propose CharacterMeet, a method to assist authors in Character construction. Authors engage in conversations with LLM-powered avatars to develop story characters. Authors are prompted to (i) describe various attributes for their character – e.g. physical description, psychological description, backstory, (ii) describe a situation in which they want to converse with the character, and then (iii) are able to interactively "chat" with that character via text or voice, and visualize that character via an avatar. By simulating dialogues with fictional personas, CharacterMeet allows authors to explore characters' backgrounds, motivations, and personalities in depth, fostering a more immersive character creation experience.

Zhang et al. (2024) propose Mathemyths, a system to assist child authors (ages 4-8) in creating stories using mathematical vocabulary. The system helps the authors learn mathematical vocabulary via collaborative narrative creation. For example, a part of the generated story could be: *. . . in the cave they find a huge pile of mystical gems, and they estimate that there are at least 100.* Mathemyths (i) prompts the author to assist them in creating narratives, (ii) builds mathematical language into the narratives via LLM-generated narrative contributions, and (iii) assists authors when they are stuck or need help. Mathemyths exemplifies how LLMs can be harnessed to create interactive educational experiences that combine creativity with curriculum goals.

Wan et al. (2024) propose Metamorpheus, a framework for recording dreams via text and image. Users input a description of each scene in their dream – literal or metaphorical – and images are generated for each scene, threaded together in the UI representation. The system offers assistance with metaphor generation/image prompting, helping users to create accurate representations of their dreams. The impact of this work is that human well-being is enhanced by emotional expression,

from which individual meaning is derived. This fusion of VLM-generated imagery and narrative highlights the potential of LLMs in therapeutic and introspective applications, where storytelling becomes a medium for personal insight and emotional processing.

Comparing these methods, CHARACTERMEET focuses exclusively on supporting authors in understanding their characters, MATHEMYTHS focuses on helping authors learn vocabulary words via story creation, and METAMORPHEUS focuses on assisting authors with expressing their dreams. More generally, these systems exemplify support for character formulation, language leveling, and expression based on a vague or emergent vision of the intended story.

Prompt-Based. Prompt-based stories are generated via static inputs from the human author(s). In this context, prompts are often single-turn directives reflecting the author's specific goals, contrasting with conversational support prompts that evolve through dialogue. In comparison to conversational systems, prompt-based systems offer minimal chat-style or brainstorming support.

Liu et al. (2024b) propose MULTIMATERIAL EXPRESSIVE ARTS THERAPY (MEAT), a method for using LLMs to enhance Expressive Arts Therapy sessions to help children and parents better express their emotions via story creation in a therapy session. First, the family creates art with traditional materials, like Legos, Play-Doh, Crayons, and more. At this stage, the art is used to create characters for later stories. Then, a picture is taken of the character and uploaded to Midjourney for refinement by the family. The character images are then physically printed out and used for physical story creation with the traditional materials again. These stories are then used to create storybooks (via Midjourney) for the children and parents to take home after the session.

3 Comparison of Datasets

We compare the datasets used in the highlighted story generation systems, as detailed in Table 1. Datasets can include story texts, comments on stories, images and their captions, and story components. Across systems, the availability and type of datasets vary widely, influencing the scope and evaluation of each method.

Several systems rely on established story text datasets – such as the Writing Prompts dataset (COLLABSTORY) and Story Commonsense dataset (MLD-EA) – to provide structured narrative inputs or benchmarks for evaluation. These datasets allow for reproducible experiments and comparative assessments.

Notably, a significant number of systems (OSOS, STORYVERSE, CHARACTERMEET, MEAT, etc.) operate without formal datasets, relying instead on user input or synthetic prompts during user studies. While this supports personalization and real-world interactivity, it limits standardization and reproducibility.

The lack of shared, diverse datasets tailored for interactive and adaptive storytelling is a major gap in current research. Expanding and standardizing datasets – especially those that integrate narrative structure, emotion, user feedback, and visual components – would greatly enhance the comparability, scalability, and realism of LLM-based storytelling systems.

4 Comparison of LLM Use

We compare the types of LLMs and their uses for story generation, as detailed in Table 1. Largely, the models are used in a prompt-based setup, leaving alternative approaches under-explored. Most systems rely on templated prompting, often with handcrafted or semi-structured inputs such as character tuples, story states, or user attributes. This reflects a trend toward controllability and interpretability, but also reveals a dependence on manual intervention and human-in-the-loop steps that may hinder scalability.

While GPT-4 and its variants dominate higher-end use cases, a number of systems (e.g., OSOS, COLLABSTORY, SWAG) demonstrate the growing capabilities of open-source models such as Llama, Gemma, and Mistral. These systems increasingly experiment with hybrid or multi-agent setups to simulate creativity (COLLABSTORY) or improve narrative coherence via iterative refinement (SWAG).

5 Comparison of Evaluations

We compare the evaluation methods used for story generation systems, as detailed in Table 1. Evaluation of LLM support for story generation includes user-focused studies of how authors and readers view the recommendations and stories generated, and automated studies assessing whether the LLM methods are generating content that meets specific

Methods	LLMs	Evaluation	Datasets	LLM Approach	Pros	Cons
OSOS (Lee et al., 2024) – *Short, vocabulary-centered stories.*	`Llama3-8B-Instruct`, `Gemma2-2B-it`, `Gemma2-9B-it`	User study with N=10 families	*No dataset.*	Templated prompt approach and reprompts with human-in-the-loop at selected steps.	• Personalized vocabulary-driven storytelling. • Human-in-the-loop enhances story relevance.	• Character visual consistency issues. • Limited to vocabulary teaching use case.
MLD-EA (Zhang and Long, 2025) – *Short, 5-sentence stories.*	`gpt-4` `StableDiffusion`	• Missing sentence detection task: P, R, F • Sentence infilling task: BLEU, ROUGE, BERTScore	Story Commonsense (Rashkin et al., 2018): approx. 5,000 5-sentence stories, use only stories with labeled emotions.	Templated prompt approach with the structured, extracted (emotion, action) character tuples as inputs.	• Improves logical and emotional coherence. • Identifies and repairs narrative gaps.	• Focused mainly on sentence-level correction. • Limited to synthetic datasets for evaluation.
PMH (Progga et al., 2024) – *Short narrative stories.*	`gpt-3.5-turbo`	Analyzed via Latent Dirichlet Allocation (Blei et al., 2003), qualitatively looking for themes in small-scale responses.	A webscraped dataset from postpartum-related forums, selecting approximately 700 narrative stories and 700 comments (Progga et al., 2023).	Templated prompt approach: combinations of co-occurence-based pairs, randomly-selected sub-theme keyword pairs (e.g. *depression*, *financial hardship*), persona, and tone.	• Focuses on real-world maternal health experiences. • Topic modeling enhances prompt design.	• LLM may refuse or hallucinate on sensitive topics. • Dataset limited in diversity.
MoPS (Ma et al., 2024) – *Short stories.*	`gpt-3.5-turbo`	• Human Evaluation • LLM-as-a-Judge	Generated premise dataset based on scraped themes, background, time, place, personas, and more.	Templated prompt approach: to control theme, background, persona, and plot modules.	• Highly diverse generated premises. • Uses sequential plot dependencies.	• Strongly-typed modules can limit creativity and diversity.
DOME (Wang et al., 2025a) – *Long stories.*	`Qwen1.5-72B-Chat`	• N-gram entropy, conflict rate • Human Evaluation: coherence, relevance, and more	DOC (Yang et al., 2023) for story premises used to generate 20 stories.	Templated prompt approach using knowledge graph tuples.	• Integrates structured KG information. • Performs well in long-context.	• Limited evaluation (20 stories). • Expensive KG module.
CollabStory (Venkatraman et al., 2025) – *Short stories.*	`Llama-2-13b-chat-hf`, `Mistral-7B-Instruct-v0.2`, `Gemma-1.1-7b-it`, `OLMo-7B-Instruct`, `Orca-2-13b`	Evaluated in terms of creativity, coherence, readability, vocabulary and sentence structure using LLM-as-a-Judge.	• Writing Prompts (Fan et al., 2018) as input • CollabStory: $> 32,000$ generated stories	Templated prompt approach: Stories generated by different orderings of LLMs with beginning, middle, and ending prompts.	• First large-scale multi-LLM collaboration dataset. • Evaluates authorship and creativity in multi-agent settings.	• Authorship attribution can be ambiguous. • Potential for conflicting narrative styles.

SWAG (Pei et al., 2024) – *Long stories.*	`Llama-2-7B`
	`Mistral-7B`
	`GPT-3.5-Turbo`
	`GPT-4-Turbo`
		• LLM-as-a-Judge: pairwise comparisons
		• Human Evaluation: pairwise comparisons of interesting-ness, surprise, coherence
			• 20,000 long LLM-generated stories
			• State-to-Action Preferences: 60,000 initial story states and next best actions from a set of 50 actions
				Supervised fine-tuning on base LLM, DPO on action discriminator LLM.	• Feedback loop improves narrative engagement.
					• Action guidance enables genre control.
						• Complexity increases with more actions.
						• Requires curated action list and fine-tuning.

CritiCS (Bae and Kim, 2024) – *Long stories.*	`gpt-3.5-turbo`	• Pairwise Human Eval. • LLM-as-a-Judge	DOC (Yang et al., 2023) for story premises.	Templated prompt approach using (generated) persona-based critics.	• Systemizes creativity. • Persona-based.	• Limited evaluation. • Focuses only on creativity.
StoryVerse (Wang et al., 2024) – *Adaptive stories.*	`gpt-4`	System demonstration via the presentation of two example stories.	*No dataset.*	Templated prompt approach using an LLM for generating character actions and narrative planning.	• Integrates author intent and emergent gameplay. • Responsive to player actions.	• Limited scalability for real-time interaction. • Evaluation based on demonstration, not user study.
CharacterMeet (Qin et al., 2024) – *Short or long stories.*	`gpt-4`	User study with N=14 users.	*No dataset.*	Templated prompt approach inputting user-provided character descriptions, backstories, and attributes to generate grounded character conversations.	• Enables deep character exploration. • Interactive, conversational interface.	• May not scale to complex narratives. • User experience highly dependent on LLM quality.
Mathemyths (Zhang et al., 2024) – *Short stories.*	`gpt-4`	User study with N=35 children ages 4-8.	*No dataset.*	Templated prompt approach using few-shot approaches for some prompts. These prompts are used for the different system modules.	• Promotes mathematical language learning. • Engaging for young children.	• Educational scope is limited (ages 4–8). • Effectiveness depends on narrative design.
Metamorpheus (Wan et al., 2024) – *Short stories.*	`gpt-3.5-turbo`	User study with N=12 users.	*No dataset.*	Templated prompt approach, inputting text and iteratively refining the output text and images.	• Supports creative and emotional self-expression. • Facilitates dream documentation.	• May produce abstract or ambiguous outputs. • Requires user effort for accurate dream recording.
MEAT (Liu et al., 2024b) – *Storybooks.*	`Midjourney`	User study with N=18 people (10 parents, 8 children, making up 7 families), supported by 4 therapists.	*No dataset.*	Templated prompt approach suggesting alternate words and phrases in a brainstorming/synonym-finding setup, and generating and refining images based on real-world constructions with materials like Play-Doh, Legos, etc.	• Blends traditional art with digital storytelling. • Family/therapist involvement enhances engagement.	• Time-intensive workflow.

Table 1: **Comparison of Systems using LLMs for Story Generation**: We compare systems in terms of the LLMs employed, the evaluation, the datasets, the LLM use, and the pros and cons.

requirements.

User studies are a common form of evaluation, used in systems like OSOS, CHARACTERMEET, MATHEMYTHS, MEAT, and METAMORPHEUS, often involving small sample sizes (ranging from 10 to 35 participants). These evaluations capture human-centered insights such as engagement, relevance, and usability, particularly for interactive or educational storytelling scenarios. However, they are often limited in scale and scope, making it difficult to generalize findings or compare systems rigorously.

Automated evaluations, on the other hand, focus on content quality through metrics like BLEU, ROUGE, and BERTScore, as seen in MLD-EA. These metrics offer reproducibility and scalability but are known to fall short in capturing creativity in narrative generation. Moreover, they often rely on synthetic or heavily curated datasets, which may not reflect real-world story complexity or user preferences. Some systems bridge these two approaches by using LLM-as-a-Judge for comparative analysis (COLLABSTORY, SWAG), combining the scalability of automated methods with closer alignment to human judgment. While promising, this approach depends on the consistency and reliability of the LLM itself as an evaluator. Further, a notable gap exists in standardized benchmarking. Additionally, evaluation setups often fail to account for longitudinal effects (e.g., user retention, narrative evolution), multimodal outputs, or collaborative authorship, despite their growing relevance in systems like STORYVERSE and SWAG.

In summary, while a variety of evaluation strategies are employed, the field would benefit from more rigorous, scalable, and standardized evaluation frameworks that integrate both human-centered and automated metrics, especially those that reflect the interactive and creative nature of story generation.

6 Discussion & Future Directions

In this section, we address some limitations of current LLM-based models in story writing and propose several potential directions for future work. Additional considerations are provided in Appendix B.

Opportunities for Multimodal Storytelling. Recent advancement in Vision-Language Models (VLMs) provides unique opportunities for generating multimodal stories. One of the key challenges is generating a sequence of coherent, contextually relevant images and texts. Many recent works (Rahman et al., 2023; Yang et al., 2024, 2025; Liu et al., 2024a) have focused on addressing this challenge. SEED-Story (Yang et al., 2024) leverages Multimodal Large Language Model (MLLM) to generate a sequence of rich and coherent narrative texts, along with images that share consistent characters and styles, given user-provided images and text as the beginning of the story. Later work StoryLLaVA (Yang et al., 2025) exploits LLaVA (Liu et al., 2023a) for generating more engaging and human-preferred narratives. In addition, Intelligent Grimm (Liu et al., 2024a) focused on open-ended storytelling by leveraging a visual-language module and a pre-trained stable diffusion model to generate unseen characters with coherent visual stories that are aligned to a given storyline.

Incorporate Constraints via Inference-Time Strategies. We propose using decoding-based constraint satisfaction methods – these methods can apply to both text stories and image consistency (Dong et al., 2024). These methods – such as constrained beam search or rule-based sampling – can enforce narrative structure, and/or character consistency without retraining. For multimodal systems, similar strategies can maintain visual coherence across scenes. This enables greater control and flexibility compared to prompt-only methods. Such approaches can enhance both the reliability and creativity of LLM-driven storytelling.

Benchmarking. No work exists yet to comprehensively evaluate the story capabilities of different LLMs. A benchmark that makes the experimental components easy to run (datasets, models, evaluation metrics) could (1) help practitioners and researchers gain an understanding of the different LLMs' performance in this area, and (2) encourage progress in this area, with enhanced resource availability.

Story-Specific Metrics. Chhun et al. (2024) propose a new large-scale automatic evaluation metric, AUTOMATIC STORY EVALUATION (ASE). This metric uses LLMs to measure a set of story-related aspects – relevance, coherence, empathy, surprise, engagement, and complexity – across a set of Likert prompts. Scores are then aggregated via correlations. This metric operates in the LLM-as-a-Judge paradigm. Future work can introduce more specific story evaluation methods.

Limitations

While LLMs have demonstrated significant potential in story generation, we now examine their limitations and ethical concerns to ensure responsible development.

Narrative Coherence and Structure. LLMs often struggle with maintaining global coherence in extended narratives. Although they can produce locally coherent text, sustaining consistent plotlines, character development, and thematic elements over longer passages remains challenging.

Contextual Understanding. LLMs may exhibit difficulties in understanding nuanced contexts, leading to inappropriate or nonsensical content generation. For instance, they might misinterpret prompts that require cultural or situational awareness, resulting in outputs that lack relevance or sensitivity.

Hallucination of Facts. A notable issue with LLMs is their propensity to hallucinate, generating information that appears plausible but is factually incorrect or unverifiable. This behavior poses risks, especially when LLMs are used in applications requiring factual accuracy, such as educational content or historical storytelling.

Acknowledgments

We thank Chengkai Liu for the discussion.

Ethical Considerations

We detail key ethical considerations with respect to using LLMs for story generation.

Authorship and Intellectual Property. The use of LLMs in story generation raises questions about authorship and ownership. LLMs trained on copyrighted materials may generate content that closely resembles existing works, leading to potential intellectual property infringements.

Authenticity and Originality. LLM-generated stories may lack the authenticity and originality inherent in human-created narratives. The reliance on patterns learned from existing texts can result in derivative works that do not offer new perspectives or insights, potentially diminishing the value of creative expression.

Impact on Creative Professions. The integration of LLMs into creative industries could disrupt traditional roles, leading to concerns about job displacement among writers and artists. While AI can augment creative processes, there is apprehension that it might replace human creativity, affecting livelihoods and the diversity of voices in storytelling.

Transparency and Accountability. The black box nature of LLMs makes it difficult to trace the reasoning behind specific outputs. This opacity challenges accountability, especially when AI-generated content causes harm. Establishing mechanisms for transparency and oversight is essential to address these concerns.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

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, Miami, Florida, USA. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2023. Storysparkqa: Expert-annotated qa pairs with real-world knowledge for children's story-based learning. *arXiv preprint arXiv:2311.09756*.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

Yee Bee Choo, Tina Abdullah, and Abdullah Mohd Nawi. 2020. Digital storytelling vs. oral storytelling: An analysis of the art of telling stories now and then. *Universal Journal of Educational Research*, 8(5):46–50.

Fatima Dayo, Ahmed Ali Memon, and Nasrullah Dharejo. 2023. Scriptwriting in the age of ai: Revolutionizing storytelling with artificial intelligence. *Journal of Media & Communication*, 4(1):24–38.

Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. A survey on llm inference-time self-improvement. *arXiv preprint arXiv:2412.14352*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Jungeun Lee, Suwon Yoon, Kyoosik Lee, Eunae Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open sesame? open salami! personalizing vocabulary assessment-intervention for children via pervasive profiling and bespoke storybook generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–32.

Zhihong Lei, Xingyu Na, Mingbin Xu, Ernest Pusateri, Christophe Van Gysel, Yuanyuan Zhang, Shiyi Han, and Zhen Huang. 2025. Contextualization of asr with llm using phonetic retrieval-based augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haotian Li, Yun Wang, and Huamin Qu. 2024. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6190–6200.

Di Liu, Hanqing Zhou, and Pengcheng An. 2024b. "when he feels cold, he goes to the seahorse"—blending generative ai into multimaterial storymaking for family expressive arts therapy. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023b. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

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, Bangkok, Thailand. Association for Computational Linguistics.

Katherine A Moreau, Kaylee Eady, Lindsey Sikora, and Tanya Horsley. 2018. Digital storytelling in health professions education: a systematic review. *BMC medical education*, 18:1–9.

Jason Ohler. 2006. The world of digital storytelling. *Educational leadership*, 63(4):44–47.

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.

Farhat Tasnim Progga, Amal Khan, and Sabirat Rubya. 2024. Large language models and personalized storytelling for postpartum wellbeing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 653–657.

Farhat Tasnim Progga, Avanthika Senthil Kumar, and Sabirat Rubya. 2023. Understanding the online social support dynamics for postpartum depression. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. Charactermeet: Supporting creative writers' entire story character construction processes through conversation with llm-powered chatbot avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Hua Xuan Qin, Guangzhi Zhu, Mingming Fan, and Pan Hui. 2025. Toward personalizable ai node graph creative writing support: Insights on preferences for generative ai features and information presentation across story writing processes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–30.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2493–2502.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Jan Niehues, and Alexander Waibel. 2025. From speech to summary: A comprehensive survey of speech summarization. *arXiv preprint arXiv:2504.08024*.

Bernard R Robin. 2008. Digital storytelling: A powerful technology tool for the 21st century classroom. *Theory into practice*, 47(3):220–228.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Maria Teleki, Xiangjue Dong, and James Caverlee. 2024. Quantifying the impact of disfluency on spoken content summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13419–13428, Torino, Italia. ELRA and ICCL.

Maria Teleki, Xiangjue Dong, Haoran Liu, and James Caverlee. 2025. Masculine defaults via gendered discourse in podcasts and large language models. In *ICWSM 2025*.

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? *arXiv preprint arXiv:2407.13248*.

Georgios Trichopoulos, Georgios Alexandridis, and George Caridakis. 2023. A survey on computational and emergent digital storytelling. *Heritage*, 6(2):1227–1263.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. CollabStory: Multi-LLM collaborative story generation and authorship analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679, Albuquerque, New Mexico. Association for Computational Linguistics.

Qian Wan, Xin Feng, Yining Bei, Zhiqi Gao, and Zhicong Lu. 2024. Metamorpheus: Interactive, affective, and creative dream narration through metaphorical visual storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025b. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Yi Wang, Qian Zhou, and David Ledo. 2024. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, pages 1–4.

Zizhen Wang, Jiangyu Pan, Duola Jin, Jingao Zhang, Jiacheng Cao, Chao Zhang, Zejian Li, Preben Hansen, Yijun Zhao, Shouqian Sun, and Xianyue Qiao. 2025c. Charactercritique: Supporting children's development of critical thinking through multi-agent interaction in story reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*.

Jing Wu and Der-Thanq Victor Chen. 2020. A systematic review of educational digital storytelling. *Computers & Education*, 147:103786.

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, Toronto, Canada. Association for Computational Linguistics.

Li Yang, Zhiding Xiao, Wenxin Huang, and Xian Zhong. 2025. Storyllava: Enhancing visual storytelling with multi-modal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3936–3951.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seedstory: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

A Paper Selection

We search in relevant top conferences in HCI and NLP, and keep relevant papers relating to story generation and Large Language Models. We look at papers from 2023-2025 to focus on the latest models, evaluation frameworks, and application studies. This helps us reflect on the emerging studies and challenges faced in automatic story generation.

The venues considered include:

- **CHI** (The ACM CHI Conference on Human Factors in Computing Systems)
- **CSCW** (The ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing)
- **ACL** (The Annual Meeting of the Association for Computational Linguistics)
- **EMNLP** (The Conference on Empirical Methods in Natural Language Processing)
- **NAACL** (The Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics)
- **COLING** (The International Conference on Computational Linguistics)
- **TACL** (Transactions of the Association for Computational Linguistics)

We also include StoryVerse (Wang et al., 2024) from The International Conference on the Foundations of Digital Games, due to its unique and relevant contribution.

B Additional Future Directions

Use Open-Sourced Models There are two ways that LLMs are currently accessible: via APIs, and locally with open-sourced models. Current work mostly uses API-based LLMs, however, there are issues with this setup: (1) There are refusal issues with API-based LLMs, as in PMH; (2) There are potential concerns with patient privacy with these models, as data is not solely kept locally; (3) In API-based systems, developers are not able to control backend changes, hindering reproducibility in the area.

Use Small Models Current work uses large LMs, but small LLMs have been shown to be comparable in quality. Methods utilizing small LLMs (e.g., distilled LLMs, etc.) could be used for practice health interventions at scale, because these models can live on a small chip on-device.

Incorporate Disfluencies Disfluencies include terms such as *uh*, and *um*, sentences that start and restart, as in *There were two dogs – I went to Target today...*, and more (Shriberg, 1994). Disfluencies are prevalent in normal spoken dialogue (Shriberg, 1994) and could be valuable for generated character dialogue. It has been shown that LLMs poorly model disfluency (Teleki et al., 2024; Retkowski et al., 2025). However, disfluencies are important for communicating emotion – and they are even carry important gender identity information (Teleki et al., 2025) – an important element of character development in stories. We propose incorporating disfluencies to express character emotions in future work.

Incorporate Discourse Features Tian et al. (2024) recently proposed a quantitative framework and dataset to benchmark and compare LLM-generated stories and human-written narratives. They show that LLMs such as GPT-4 and Claude cannot generate narratives comparable to human-level storytelling on certain aspects such as story arc development, turning points, and affective measures (arousal and valence). Moreover, these LLMs exhibit limited understanding of these discourse-level features and thus generate rather uniform structural patterns, with inadequate reasoning and a shallower portrayal of emotional dynamics. Although integrating such discourse-level elements explicitly helps create more diverse and engaging narratives, current models still cannot sufficiently capture the full complexity and emotional depth of human storytelling, especially when handling more dark and negative plot lines. A future direction in this regard would be to develop nuanced ways of analyzing discourse in narratives and to develop models that are more aware of these features.

Fused Embedding Approaches A pre-LLM[1] work, CHARGRID (Hong et al., 2023), takes a fused embedding approach, designing an approach to include a specialized character embedding. Character consistency is often an issue in generations – e.g., using the name *David* to refer to the same character across multiple input image scenes. CHARGRID features a specialized character embedding that is input to the transformer to assist in creating character-faithful generations. This embedding is concatenated with the other embeddings in the architecture. Hence, CHARGRID successfully maintains faithfulness to characters throughout the generations. This type of embedding-based methodology should also be explored in the LLM era, given that there is a vast literature of embedding-based work (Liu et al., 2023b; Cao et al., 2024; Wang et al., 2025b; Lei et al., 2025). These types of methods can be specifically designed for story generation.

C Desired Features in Creative Support Tools

STORYNODE (Qin et al., 2025) explores potential features to assist authors with story writing via small-scale human feedback with a formative study (N=12), a user study (N=14), and an external study (N=19). In this work, they explore features such as: chat with various personas for manuscript feedback, story modification via suggested prompts, generation of music/image, and plot event graph conversion. They find that users find chatbot-simulated conversations with characters unhelpful and unrealisic.

D A Related Task – Story QA

In a related direction, CHARACTERCRITIQUE (Wang et al., 2025c) explores how LLMs can engage children and their parents in question-answer dialogues tied to the story they are reading. Using GPT-4o, multiple AI agents can role-play as either story characters or user-designed personas and interact with children to foster analytical and cognitive skills. While user studies show promising results, current LLMs still struggle to generate compelling visual scenes as well as accurately interpret children's verbal and non-verbal responses. Another system, STORYSPARKQA (Chen et al., 2023) also focuses on QA for children's stories, highlighting that their dataset construction method can help to "capture the nuances of how education experts think when conducting interactive story reading activities." They release a dataset of annotated QA pairs for this task. Future work can build on these contributions and improve the interpretation of children's responses and incorporate specialized QA into adaptive storytelling approaches.