A Survey on LLMs for Story Generation

Maria Teleki[1], Xiangjue Dong[1], Peter Carragher[2], Vedangi Bengali[1], Tian Liu[1], Haoran Liu[1], Sai Tejas Janjur[1], Thomas Docog[1], Stephanie Birkelbach[1], Oliver Grabner[1], Cong Wang[1], Ting Liu[1], Yin Zhang[1], Frank Shipman[1], James Caverlee[1]

[1]Texas A&M University; [2]Carnegie Mellon University

Extended Version. This paper extends our EMNLP Findings 2025 survey with additional content and updated coverage. The original version is available here: `https://aclanthology.org/2025.findings-emnlp.750/`

Abstract

Methods[1] 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 systems based on their methodology, datasets, generated story types, evaluation practices, and model usage, with special attention to the structured representations – knowledge graphs, narrative arcs, event schemas, planner–generator frameworks, and memory models – that increasingly shape controllable, explainable story generation. Building on these insights, we outline key directions for future research, including story personalization, deeper integration of structured knowledge, expansion of planner-generator frameworks, and more expressive modeling of human emotion, values, and interaction styles within narrative systems.

Keywords: Story Generation, LLM, AI, Narrative, Storytelling, Survey

1 Introduction

Highly capable LLMs like ChatGPT, Llama, and more [1, 17] open up possibilities to rethink and re-formulate the static, existing ways of storytelling [9]. For example, **with LLMs, stories can be interactive [70] and personalized [20]**, 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 [70], improving childcare quality and training health professionals in the healthcare industry [31], improving teaching methods in education [35, 53], and improving movie script development in the entertainment industry [10].

Despite the appealing capabilities of LLMs, they struggle to maintain global narrative coherence, track evolving story states, or enforce long-range constraints across characters, events, and themes. These limitations stem from the fundamentally unstructured nature of standard next-token prediction: without explicit guidance, LLMs drift, contradict earlier details, or fail to sustain plot arcs over extended text. Recent work demonstrates that structured representations – such as narrative plans, discourse arcs, knowledge graphs, event schemas, persona graphs, and memory or world models – provide the scaffolding needed to overcome these weaknesses, and we survey them in this work. **By introducing structure, systems can anchor generation in stable semantic relationships, enforce temporal or causal constraints, and support consistent character development.** Moreover, structured approaches enable explanations, controllability, and interaction modalities that LLMs alone cannot reliably offer.

While there is a vast literature of work on digital storytelling [35, 61, 73] and the use of traditional language models – i.e., pre-LLM[1] – for story generation [2, 14], there is a gap in the literature in surveying the use of LLMs in story generation. The closest prior work, Li et al. [22], 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[2] 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, the use of inference-time methods for effectively controlling LLM outputs [11, 72]. We also emphasize the growing importance of structured representations for improving narrative control and coherence (e.g. [47, 68]). We make the following contributions:

[1]*Preprint, Under Review.*

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

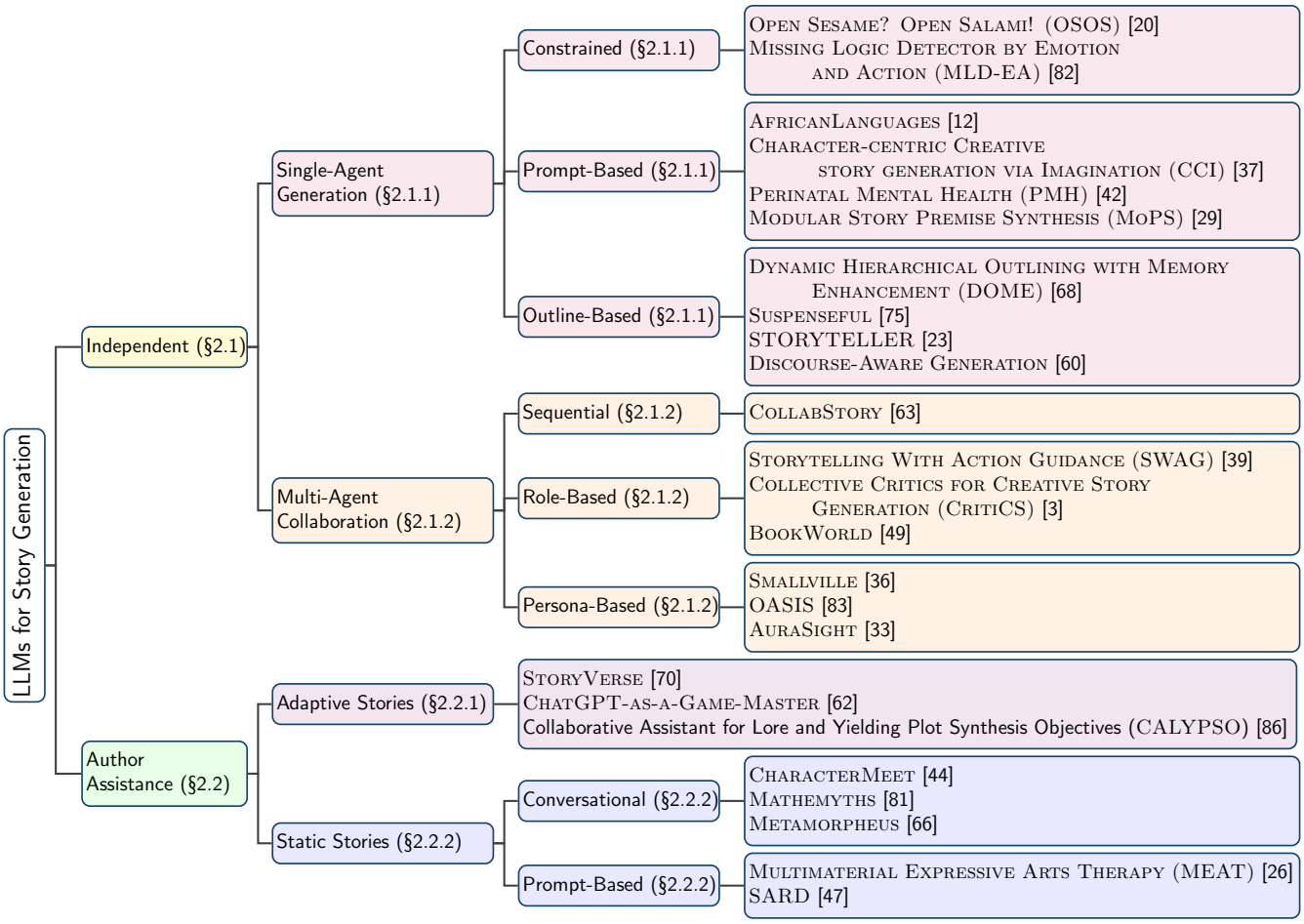

Figure 1: Taxonomy of LLMs for Story Generation. *Note that, for clarity, unnamed frameworks are assigned descriptive labels reflecting their key contributions.*

- We introduce a novel **Taxonomy of LLMs for Story Generation** (Figure 1, §2), categorizing 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 1. 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. Independent story-generation methods signify a broader shift toward modes of narrative production in which human authorship is no longer a strict requirement. By positioning the model as the primary author, these approaches enable the creation of narratives that are sufficient for domains where literary nuance is less central than scalability, adaptability, or responsiveness – e.g., educational settings.

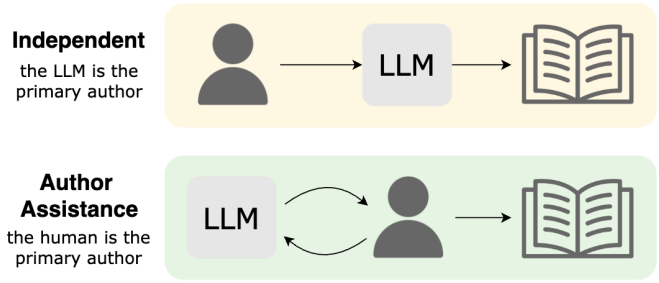

Figure 2: **Major Taxonomy Categories**: In our taxonomy (Figure 1), we first categorize works based on primary authorship.

In our view, the objective of these methods should not be to rival human writers, but to furnish stories that meet the functional demands of such contexts. Most of the works in this section therefore focus on probing the fundamental extent to which LLMs can independently produce coherent stories, serving as an initial step in this emerging direction.

2.1.1 Single-Agent Generation

Single-Agent Generation methods include the use of one agent for generation, in contrast to multi-agent paradigms. We divide the approaches in this category into methods that methods differ mainly in their method of control imposed on the model: constrained, prompt-based, and outline-based.

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. [20] 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: the Profiler, the Extractor, and the Generator. The Profiler takes audio input from the child's home and converts the audio to diarized text. The intuition behind this design decision is that children are exposed to different sets of words based on their home environments – e.g. one family may allow their children to watch *Bluey*, a show about a dog, while another may want them to watch *Cyberchase*, a show about kids solving math problems – and hence, the two kids should have different words prioritized for them by the system so they can be successful in their home environments. 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: 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", a check on the image generations, and a check on the final story. With this system, the top k words serve as structure, as well as the overall pipeline of the system constraining the generation process.

Zhang and Long [82] 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

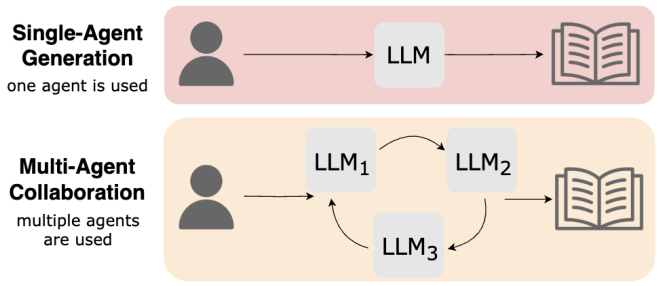

Figure 3: **Independent Generation Categories** (§2.1)

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.

Essuman and Buys [12] investigate whether small LMs can generate coherent children's stories in isiZulu and Yoruba when trained on synthetic data (AfricanLanguages). They construct six datasets through machine translation of TinyStories or direct multilingual prompting, and then train corresponding models from scratch. The models are evaluated on perplexity, lexical and semantic diversity, and Gemini-based qualitative scoring. Results show that models trained on larger translated datasets generalize best, while prompt-generated data yields more creative stories but weaker generalization. Additionally, models trained on synthetic data can produce grammatical and coherent stories, but the creativity and plot depth of these stories falls short. This work illustrates that synthetic corpora can effectively support story generation models in low-resource languages, making independent LM development feasible for African languages.

The CCI (Character-centric Creative story generation via Imagination) framework [37] targets long-form stories that are more creative, diverse and strongly character-driven. It augments a a standard plan-and-draft LLM pipeline with an Image-guided Imagination module (IG) and Multi-Writer module (MW) for explicit persona modeling. IG first generates images from brief textual prompts (using DALL-E 3) and then uses GPT-4o to produce detailed textual descriptions of the depicted characters, backgrounds, and key events. This yields more vivid characters, background and key plot events. Next, a Specification step expands and links the protagonist to the main plot. MW (also using gpt-4o) then creates and reranks multiple persona snippets, and injects them throughout the narrative, keeping the protagonist's character consistent and more aligned with the main plot. This multi-stage, template-prompted pipeline is model-agnostic and improves story diversity and character concreteness. However, the persona modeling involves only a subset of traits, and the framework is mainly evaluated with a single backbone and relatively narrow baselines. Overall, CCI illustrates how adding multimodular "imagination" and character-centric personas can enhance diversity and creativity of LLM-generated stories.

Progga et al. [42] 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. [29] 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. Outline-based methods incorporate an explicit planning stage, typically leveraging structured representations of narrative progression or established theories of story arc. These approaches guide generation by first scaffolding the plot – often through hierarchical outlines, event sequences, knowledge graphs, or discourse-level structures – and then expanding these plans into full narratives, or by using iterative planner-generator patterns. By exploiting known models of narrative development, including formulations such as Vonnegut's plot shapes [65], outline-based systems aim to improve global coherence, control plot evolution, and reduce the drifting behaviors common in unconstrained generation.

Wang et al. [68] propose 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, 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, 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.

Xie and Riedl [75] propose an iterative outline-based planning approach for Suspenseful long-form story generation with LLMs. Their outlines are grounded in a psychological theory of suspense in which a protagonist faces an impending unfavorable outcome while the number of viable escape paths shrinks. The system alternates between prompting the LLM to propose protagonist plans and introducing adversarial story conditions that cause those plans to fail, iteratively constructing a suspenseful action–failure outline. This outline is then expanded by the LLM into a multi-chapter narrative via a separate elaboration stage. Human evaluations show that this method produces stories with higher perceived suspense, novelty, enjoyment, and logical sense than direct prompting.

Li et al. [24] propose STORYTELLER, which uses knowledge graphs (KGs) to represent the subject-verb-object storyline and narrative entities, respectively. These KGs interact throughout generation to update plot points adaptively, enabling coherent narrative structure and fine-grained control over plot progression. This design preserves logical flow, strengthens thematic consistency, and enhances narrative depth.

Tian et al. [60] propose Discourse-Aware Generation, which incorporates discourse features into story generation – specifically, story arcs, turning points, and affective dimensions (e.g. arousal, valence). These discourse features are selected because they cover macro-, meso-, and micro-levels of the narrative. Interestingly, the story arcs are derived from a lecture by Kurt Vonnegut [65], where the x-axis is time, and the y-axis is *good/bad fortune*. The authors find that incorporating discourse-aware features into generation via planning improves performance.

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. In our view, this paradigm offers a distinct advantage: the heterogeneity of models can lead to richer variation of style and content than a single model can reliably produce on its own, making multi-agent collaboration a promising direction for enhancing narrative diversity and exploring emergent creative dynamics.

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. [63] 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. We also view this paradigm as offering an additional advantage: role differentiation makes it possible to incorporate domain-specialized personas as

expert contributors to the narrative. For example, an agent instantiated as an archaeologist could provide grounded domain knowledge for an Indiana Jones–style adventure, while other expert agents might supply historical or scientific knowledge. Such specialization allows Role-Based systems to draw on expertise, enriching story plausibility and depth without sacrificing centralized narrative control.

Pei et al. [39] 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 [3] 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.

BookWorld [49] proposes a framework for constructing and simulating book-based multi-agent societies that simulates story progression in fictional worlds and facilitates story creation. The system extracts structured character profiles, relationships, and worldview settings directly from source books, which are then used to initialize "Role Agents" possessing static and dynamic attributes, actions, and short-term and long-term memory, alongside a "World Agent" responsible for environmental responses, and event generation and updates. The system operates through scene-based interaction cycles, the logs of which are subsequently rephrased by LLMs into cohesive, novel-style narratives.

Persona-Based. Scenario generation is an alternate use-case for generative storytelling. Here, the environment is controlled by a social (character) simulation via persona-instantiated LLMs, and the story that unfolds emerges from the interactions of various LLM agents as they respond to changes in their environment based on their assigned personas.

Park et al. [36] create a two-dimensional sandbox environment for agents to interact, called Smallville. In it, they instantiate 25 agents using the GPT-4 API, and have each represent a different character in the world of Smallville. Park et al. [36] focus on the emergent behavior of these agents as they naturally form friendships, and begin to plan events together. This paradigm has inspired a new wave of research applying LLMs to traditional agent-based simulations [16, 33, 64].

OASIS [80] introduces a generalizable and scalable social media simulator built upon LLM agents. It incorporates key elements of real platforms—dynamic social networks, continuously updated content streams, diverse user actions (e.g., posting, commenting, following), and practical recommendation systems (interest-based and hot-score ranking). The system consists of five components: an Environment Server for global platform state, a Recommender System module for content ranking, an Agent Module with LLM-driven behavioral reasoning, a Time Engine for orchestrating activity over simulated time, and a Scalable Inferencer enabling efficient large-scale execution. Crucially, OASIS can simulate up to one million agents, far exceeding the scale of prior LLM-based simulations. This allows researchers to investigate complex emergent behaviors – such as information diffusion, opinion dynamics, polarization, and herd effects – in a controlled and reproducible setting that more closely reflects real-world social platforms.

Ng et al. [33] extend scenario generation to synthetic social networks with AuraSight. The authors first establish a scaffolding for their scenario by setting up groups of social LLM agents that represent different factions within their synthetic social network. Groups are formed based on the agent's personas, by assigning them demographic attributes, relationships with other agents, as well as a collective identity or ideology. Ng et al. [33] then describe a series of provocative incidents, events that are designed to lead the various key actors from different groups into conflict with each other.

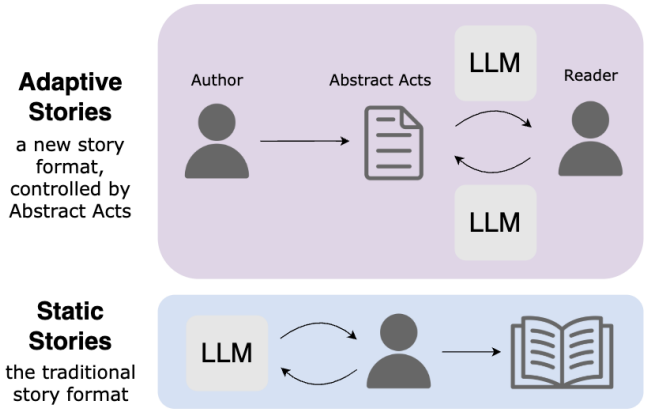

Figure 4: **Author Assistance Categories** (§2.2)

2.2 Author Assistance

In contrast to independent story generation, where LLMs act as the primary author, author assistance 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. In our view, these methods reflect an expanding ceiling on human creativity: rather than replacing the author, they extend the author's expressive range by offering flexible, adaptive support throughout the writing process. The systems in this paradigm are therefore best understood as tools designed to serve the author – augmenting ideation, structure, and refinement.

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. These systems are particularly compelling because they revisit a long-standing but previously limited narrative form. Earlier attempts at branching or choice-based storytelling – such as "Choose Your Own Adventure" books – were constrained by rigid, pre-authored paths, independent of specific knowledge of the reader. In contrast, LLM-driven adaptive stories can generate **personalized** narrative trajectories in real time, enabling far more natural integration of reader choices. This allows for richer expressions of reader agency and interactive experiences that feel less like navigating predefined branches and more like participating in a dynamically unfolding story.

In a particularly unique setup, Wang et al. [70] 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.

However, in comparison with synthetic social media scenarios [33, 36], the StoryVerse plot plan is still more thoroughly defined. This is in part due to the challenges inherent in attempting to simulate the effects of a wide range of player interactions on the game's environmental state. The more degrees of freedom the player has to manipulate the environment, the more difficult it becomes to track the environmental state. Game designers must strike a balance between maintaining the illusion of player agency and crafting a coherent plot structure.

This balance lies at the heart of the popular table-top role-playing game *Dungeons & Dragons* (D&D). In D&D, players each take on a different character persona, forming a group that will overcome various challenges together. Designing these challenges, and the world in which the challenges are placed, is the Game Master (GM). The richness of interactions during D&D gameplay has led to D&D dialogue datasets becoming benchmark NLP datasets and challenge problems in NLP [5, 85]. Zhou et al. [85] presents LLM guidance as one such problem, where LLMs are tasked with subtly guiding D&D players along a predefined plot path, without breaking the illusion of player agency.

D&D is typically described as a collaborative storytelling experience, where the storyline progresses with the back-and-forth communication between the players and the GM. However, the constant back-and-forth style of D&D means that, try as they might, the GM likely cannot predict or prepare for the players' actions. Instead, they will likely be forced to improvise. Motivated by the cognitive effort of this task, Santiago et al. [54] proposes several ways in which LLM story generation can aid GMs. To this end, Zhu et al. [86] introduces CALYPSO, a system of GPT-3-based tools that support DMs, both in their preparation and in real-time during the game itself. Finally, Triyason [62] go one step further—they employ CHATGPT-AS-A-GAME-MASTER in a series of pilot experiments with real players. Their findings suggest that LLMs can be successful in end-to-end story generation and delivery in real-world game management scenarios.

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 an unchanging fixed string. These stories align with the longstanding tradition of fixed-form narrative, but the supporting tools represent a meaningful evolution in how such stories can be produced. These systems can help emerging authors produce higher-quality work and enable experienced writers to develop material more efficiently. In this sense, static-story assistance does not alter the fundamental nature of the final product, but it can substantially ease the creative process and broaden the range of authors who can realize their narrative intentions.

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. [44] 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. [81] 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. [66] 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.

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.

Methods	LLMs	Evaluation	Datasets	LLM Approach	Pros	Cons
OSOS [20] – *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.

Continued on next page

Methods	LLMs	Evaluation	Datasets	LLM Approach	Pros	Cons
MLD-EA [82] – *Short, 5-sentence stories.*	`gpt-4` `StableDiffusion`	• Missing sentence detection task: P, R, F • Sentence infilling task: BLEU, ROUGE, BERTScore	Story Commonsense [50]: 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.
African Languages [12] – *Short stories.*	`gpt-2` `AfroLlama-V1` `SeamlessM4T-V2`	• Perplexity, Type-Token Ratio, Semantic Similarity • LLM-as-a-judge: grammar, coherence, plot, creativity	• 250k translated TinyStories (isiZulu/Yoruba) • 10k translated subsets • 10k stories generated via multilingual prompting	Single-agent generation where models are trained entirely on synthetic corpora created via translation or prompting.	• Coherent and grammatical story generation. • Demonstrates feasibility of low-resource LM development using only synthetic data.	• Creativity and plot complexity remain limited. • Prompt-generated datasets show poor generalization.
CCI [37] – *Long stories.*	`DALL-E-3`, `gpt-3.5 -turbo`, `gpt-4o`	• Human Evaluation • LLM-as-a-Judge	ReedsyPrompts [38] dataset used for training the Multi-writer module; no new story dataset introduced.	Templated prompt approach: with modular stage-specific prompts for imagination, specification, planning, drafting, and persona control.	• Multimodal (image+text) imagination improves story creativity and diversity. • Model-agnostic plug-in modules that do not require LLM fine-tuning.	• Only a subset of character traits are incorporated in Persona modeling. • Evaluated mainly with one backbone framework, so generality is unclear.
PMH [42] – *Short narrative stories.*	`gpt-3.5-turbo`	Analyzed via Latent Dirichlet Allocation [4], qualitatively looking for themes in small-scale responses.	A webscraped dataset from postpartum-related forums, selecting approximately 700 narrative stories and 700 comments [43].	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 [29] – *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 [68] – *Long stories.*	`Qwen1.5-72B-Chat`	• N-gram entropy, conflict rate • Human Evaluation: coherence, relevance, and more	DOC [77] 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.
Suspenseful [75] – *Long stories.*	`gpt-3.5 -turbo`, `Llama-2-13b -chat`	• Human Evaluation: pairwise comparisons of suspense, novelty, enjoyment, logic, naturalness.	*No dataset.*	Templated prompt approach: background setup, iterative action–event outline, then chapter elaboration.	• Improves human perceived suspense and novelty.	• Multi-stage prompting . • Limited evaluation.
STORYTELLER [24] – *Adaptive stories.*	`gpt-4o`	Both LLM-as-a-Judge and pilot user study.	WRITING-PROMPTS [13]	Conversational approach using ChatGPT to interact directly with players and adapt to their actions.	• Graph guidance enhances coherence and relevance.	• Reliance on KG-guided structure may unintentionally restrict creative phrasing and stylistic diversity.
Discourse-Aware Generation [60] – *Short stories.*	`gpt-3.5`, `gpt-4`, `Llama3-8B`, Gemini Pro, `Claude3`	• Human evaluation of suspense, emotion provocation, and overall preference. • Turning point identification.	Narrative Discourse [60]	Templated prompt approach: First prompt to generate an outline, then expanded to generate full story; compared to self-annotation of turning points for generation.	• Reasoning about *turning points* reduces plot holes, enhances suspense + emotional aspects.	• Highly reliant on annotated data. • Small dataset (approx. 800 samples).
CollabStory [63] – *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 [13] 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 [39] – *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 [3] – *Long stories.*	`gpt-3.5 -turbo`	• Pairwise Human Eval. • LLM-as-a-Judge	DOC [77] for story premises.	Templated prompt approach using (generated) persona-based critics.	• Systemizes creativity. • Persona-based.	• Limited evaluation. • Focuses only on creativity.

Continued on next page

Methods	LLMs	Evaluation	Datasets	LLM Approach	Pros	Cons
BookWorld [49] – *Novel-style narratives.*	`gpt-4o`, `gpt-4o-mini`, `gemini-2`, `qwen-plus`, `deepseek-v3`, `llama-3.3-70b`, `qwen2.5-72b`	Pairwise Win Rate (judged by gpt-4o) vs. Direct Generation & HoLLMwood, Human agreement	16 Novels: 10 English, 6 Chinese.	Templated prompt approach: Multi-Agent System featuring "Role Agents" (RAG-based memory of profiles/text excerpts) and a "World Agent" (manages environment/events).	• Higher immersion and setting fidelity. • Supports both autonomous ("Free") and user-guided ("Script") modes.	• Writing quality sometimes lags behind specialized role-playing frameworks. • Performance drops significantly on models with weaker bilingual capabilities.
Smallville [36] – *Persona-based stories.*	`gpt-4`	Quantitative surveys measured believability of LLM agent actions and rationale during simulation.	*No dataset.*	Templated prompt approach to condition LLM generation on personas from the social simulation.	• Emergent social behaviors form stories organically.	• Output is a stream of social network interactions, not a single text.
OASIS [80] – *Social posts.*	`Llama3-8b -instruct`	Information spreading, group polarization, herd effect	198 real-world propagation instances (X), 116,932 real Reddit comments, 21,919 counterfactual posts	Templated prompt approach Agent-Based Model (ABM)	• Highly scalable (up to 1M agents). • Generalizable to different platforms (X and Reddit).	• Simulating millions of agents still takes several days.
AuraSight [33] – *Persona-based stories.*	`gpt-4`	The synthetic social network formed from LLM agent interactions was validated against social network theory.	*No dataset.*	Templated prompt approach to condition LLM generation on personas from the social simulation.	• Scales to interactions between 1000's of agents.	• Output is a stream of social network interactions, not a single text.
StoryVerse [70] – *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.
ChatGPT-as-a-Gamemaster [62] – *Adaptive stories.*	`ChatGPT`	Pilot user study with 4 inexperienced players.	*No dataset.*	Conversational approach using ChatGPT to interact directly with players and adapt to their actions.	• High degree of adaptability according to user feedback.	• Reliance on ChatGPT led to low player immersion.
CALYPSO [86] – *Adaptive stories.*	`gpt-3`	Qualitative user study with 71 players.	*No dataset.*	Templated prompt approach using an LLM for generating D&D encounters, and conversational tools for encounter refinement.	• Handles rule based nature of D&D mechanics well.	• Limited by diversity and creativity of generated responses.
Character Meet [44] – *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 [81] – *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 [66] – *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 [26] – *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.
SARD [47] – *Multi-chapter stories.*	`gpt-4`	User study with N=5 graduate creative writing students.	*No dataset.*	Templated prompt approach mapping node-based storyboards into chapter-generation and summarization prompts.	• Interface helps story mental modeling. • Helps novice writers elaborate upon simple plots.	• Hidden prompts and limited lexical diversity in generated content.

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. [26] 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.

Radwan et al. [47] propose SARD, a co-creative authoring tool for multi-chapter stories built around a node-based storyboard canvas. Authors construct graphs of characters, actions, and relationships with images to represent characters and scenes. These images are then described via GPT-4 and, together with the structured storyboard and an explicit ordering of events, are converted into templated prompts for GPT-4 to generate individual chapters and brief internal summaries that maintain cross-chapter coherence. The authors find that this visual interface helps writers structure their narratives, but it can become cluttered and cognitively demanding as stories grow in complexity, and hidden prompts limit authors' direct control over the generated text.

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. While these datasets fail to capture the richer interactive dynamics surfaced through user studies, they do allow for reproducible experiments and comparative assessments.

In contrast, 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.

More broadly, these divergent evaluation strategies reflect long-standing differences between HCI and NLP research traditions. Because both offer complementary insights, we recommend adopting hybrid approaches that incorporate standardized datasets alongside user-centered evaluations.

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 requirements.

Yang and Jin [76] advance automated evaluation by exploring effective methods for long-story assessment and proposing structured, multi-aspect criteria. They compare aggregation-based, incrementally updated, and summary-based strategies, concluding that aggregation- and summary-based approaches offer stronger performance in terms of detail assessment and efficiency, respectively. Building on these insights, they introduce NovelCritique, an efficient 8B model that employs a summary-based reviewing framework to score stories across predefined dimensions.

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.

6.1 Addressing Evaluation Limitations

Develop Unified Benchmarks. 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.

Develop Story-Specific Metrics. Chhun et al. [8] 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. Additionally, metrics can be developed based on *structural analysis*, which entails computing measures over consecutive sections of the text to determine attributes such as composition and the Flow of Information (FoL). For example, Philipp et al. [40] presents a novel method for revealing information density and subtext. By computing suprisal values over a Chekhovian short story and contrasting these with predictions from the theory of Uniform Information Density, they can determine sections in a text where information density is higher than expected—sections that may contain subtext. They consider the effect that enriching the text with syntactic glosses generated by LLMs and topic models has on these surprisal values. In developing interpretable information theoretic measures, this research represents a step towards subtext understanding and inference in LLMs. This is a promising avenue for future work, and research in this area may benefit from the wealth of related research on sarcasm and irony detection [41]. Future work can introduce more specific story evaluation methods.

6.2 Addressing Structural Limitations

Incorporate Discourse Features via Structured Knowledge. In Discourse-Aware Generation [60], Tian et al. 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, perhaps integrating stronger representational tools such as knowledge graphs similarly to Li et al. [23].

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 [11]. 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.

Incorporate Fused Embedding Approaches. A pre-LLM[1] work, CharGrid [18], 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 [6, 21, 28, 69]. These types of methods can be specifically designed for story generation.

Expand Planner-Generator Approaches. Planner-generator architectures have emerged as one of the most effective strategies for controllable long-form generation, particularly within outline-based systems (§2.1.1). Their strength lies in decomposing narrative construction into structured planning followed by focused generation, allowing models to circumvent the weaknesses of long-context decoding and maintain coherence across large spans of text. This staged workflow allows for greater control, enabling systems to enforce high-level plot structure, thematic progression, or character arcs before any prose is produced.

Looking forward, other works can also integrate the planner-generator paradigm in other settings. Multi-agent frameworks (§2.1.2) are a natural fit, because planners can be distributed across specialized agents (e.g., plot architects, domain specialists, etc.), which then feed into a generator agent responsible for prose realization. Similarly, Author Assistance tools (§2.2) can adopt planner–generator cycles to provide writers with iterative structure. Human authors or groups of LLM agents could collaboratively modify plans, and rely on generators for generating the prose. Overall, continued expansion of planner-generator methods offers a promising path toward more controllable, interpretable, and structurally robust narrative systems.

Focus on 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 [25, 48, 78, 79] have focused on addressing this challenge. SEED-Story [79] 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 Story-LLaVA [78] exploits LLaVA [27] for generating more engaging and human-preferred narratives. In addition, Intelligent Grimm [25] 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. Recent work [74] proposes a Visual Story Ideation task that aims to automatically generate multiple storylines from a collection of visual assets (e.g., images and video frames). Specifically, they leverage MLLMs to first extract visual features and textual information from the given visual assets, and then select candidate assets using LLM-created story graphs, followed by multiple story generation with MLLMs.

Design Spoken Story Frameworks. As voice-first interaction becomes more common, users will increasingly want to *speak with their stories* rather than only read them. Spoken narratives impose constraints that differ from written text: they rely on prosody, pacing, repetition, and real-time turn-taking, and must respond fluidly to interruptions or shifts in user intent. Current LLM story systems are optimized for written output and do not account for these dynamics. Future frameworks can incorporate prosody, dialogue-focused narrative structures, and duplex-style responsiveness [19, 30, 67]. This would better support interactive, performed stories and extend LLM storytelling into voice-driven settings such as education, home devices, and participatory audio narratives.

6.3 Addressing Tooling Limitations

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.

6.4 Addressing Human Modeling Limitations

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 [56]. Disfluencies are prevalent in normal spoken dialogue [56] and could be valuable for generated character dialogue. It has been shown that LLMs poorly model disfluency [52, 58]. However, disfluencies are important for communicating emotion – and they are even carry important gender identity information [59] – an important element of character development in stories. We propose incorporating disfluencies to express character emotions in future work.

Incorporate Human Value & Emotion Modeling. Current story generation systems generally simulate characters via persona instantiations [15], neglecting to incorporate richer representations of the human psyche. Recent advances in modeling human values [57] and emotions[55] are available to story generation system developers, and offer richer frameworks for encoding motivations, emotional dynamics, and internal conflicts – e.g. *Should I downplay my achievements in a competition to let my friend shine?* Integrating these models would allow narrative systems to simulate characters whose decisions are grounded in more realistic psychological processes, to create stories with greater character depth.

Incorporate Personalization. LLM personalization is still emerging, and offers interesting implications for the adaptive story paradigm (§2.2.1). Approaches to personalization can range from lightweight techniques such as query rewriting to far more sophisticated methods inspired by recent hybrid neural recommender-system frameworks, which learn user-specific preference representations and personalized models [34, 46, 87]. Bringing personalization into generative storytelling opens a promising frontier, enabling narratives that respond more faithfully to individual readers' histories, preferences, and interaction patterns.

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.

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.

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.

Impact on Creative Professions. The integration of LLMs into creative industries could disrupt traditional roles, leading to concerns about job loss 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. Yet AI also opens the door to genuinely new forms of narrative creation – interactive, adaptive, and more dynamic than what has been previously possible. The systems discussed in this survey illustrate how LLMs can broaden, rather than narrow, the creative landscape. We are excited about what the future holds, and encourage authors to explore how these tools might enrich their own storytelling practices.

Acknowledgments

We thank Chengkai Liu for the discussion.

Disclosure AI tools were used to assist with interpreting, organizing, and refining writing ideas. All outputs were reviewed, verified, and modified by the authors.

References

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

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

[3] 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 18784–18819. doi:10.18653/v1/2024.emnlp-main.1046

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

[5] Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Abu Dhabi, United Arab Emirates, 2022-12), Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 9379–9393. doi:10.18653/v1/2022.emnlp-main.637

[6] 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 (2024), 49519–49551.

[7] 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* (2023).

[8] 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 (2024), 1122–1142. doi:10.1162/tacl_a_00689

[9] 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 (2020), 46–50.

[10] 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 (2023), 24–38.

[11] Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. A Survey on LLM Inference-Time Self-Improvement. *arXiv preprint arXiv:2412.14352* (2024).

[12] Catherine Nana Nyaah Essuman and Jan Buys. 2025. Story Generation with Large Language Models for African Languages. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, Constantine Lignos, Idris Abdulmumin, and David Adelani (Eds.). Association for Computational Linguistics, Vienna, Austria, 115–125. doi:10.18653/v1/2025.africanlp-1.16

[13] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[14] 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 (2023), 14361–14397.

[15] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling Synthetic Data Creation with 1,000,000,000 Personas. arXiv:2406.20094 [cs.CL] https://arxiv.org/abs/2406.20094

[16] Navid Ghaffarzadegan, Peiran Wang, and Peter Brown. 2024. Generative Agent-Based Models: Simulating Human-Like Social Behaviors Using Large Language Models. *Journal of Computational Social Science* 6 (2024), 1–15.

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[18] 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 (2023), 565–581. doi:10.1162/tacl_a_00553

[19] Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. 2025. Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model. *arXiv preprint arXiv:2505.15670* (2025).

[20] 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*. 1–32.

[21] 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)*. IEEE, 1–5.

[22] 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*. 1–19.

[23] Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, et al. 2025. STORYTELLER: An Enhanced Plot-Planning Framework for Coherent and Cohesive Story Generation. *arXiv preprint arXiv:2506.02347* (2025).

[24] Jiaming Li, Yukun Chen, Ziqiang Liu, Minghuan Tan, Lei Zhang, Yunshui Li, Run Luo, Longze Chen, Jing Luo, Ahmadreza Argha, Hamid Alinejad-Rokny, Wei Zhou, and Min Yang. 2025. STORYTELLER: An Enhanced Plot-Planning Framework for Coherent and Cohesive Story Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 20818–20846. doi:10.18653/v1/2025.findings-acl.1071

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

[26] Di Liu, Hanqing Zhou, and Pengcheng An. 2024. "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*. 1–21.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

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

[29] 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)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 2135–2169. doi:10.18653/v1/2024.acl-long.117

[30] Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025. Language model can listen while speaking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 24831–24839.

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

[32] Sara Nabhani, Khalid Al Khatib, Federico Pianzola, and Malvina Nissim. 2025. Storytelling in Argumentative Discussions: Exploring the Use of Narratives in ChangeMyView. In *Proceedings of the 12th Argument mining Workshop*, Elena Chistova, Philipp Cimiano, Shohreh Haddadan, Gabriella Lapesa, and Ramon Ruiz-Dolz (Eds.). Association for Computational Linguistics, Vienna, Austria, 217–227. doi:10.18653/v1/2025.argmining-1.21

[33] Lynnette Hui Xian Ng, Bianca N. Y. Kang, and Kathleen M. Carley. 2025. AuraSight: Generating Realistic Social Media Data. arXiv:2509.08927 [cs] doi:10.48550/arXiv.2509.08927

[34] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2025. User-llm: Efficient llm contextualization with user embeddings. In *Companion Proceedings of the ACM on Web Conference 2025*. 1219–1223.

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

[36] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs] http://arxiv.org/abs/2304.03442

[37] 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*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 1598–1645. doi:10.18653/v1/2025.findings-acl.82

[38] Kyeongman Park, Nakyeong Yang, and Kyomin Jung. 2024. Longstory: Coherent, complete and length controlled long story generation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 184–196.

[39] 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*. 14086–14106.

[40] J Nathanael Philipp, Olav Mueller-Reichau, Matthias Irmer, Michael Richter, and Max Kölbl. 2025. Can information theory unravel the subtext in a Chekhovian short story?. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*. 84–90.

[41] Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications* 32, 23 (2020), 17309–17320.

[42] 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*. 653–657.

[43] 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*. 1–17.

[44] 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*. 1–19.

[45] 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*. 1–30.

[46] Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025. Latent inter-user difference modeling for llm personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 10610–10628.

[47] Ahmed Y. Radwan, Khaled M. Alasmari, Omar A. Abdulbagi, and Emad A. Alghamdi. 2024. SARD: A Human-AI Collaborative Story Generation. arXiv:2403.01575 [cs.HC] https://arxiv.org/abs/2403.01575

[48] 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*. 2493–2502.

[49] 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)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 15898–15912. doi:10.18653/v1/2025.acl-long.773

[50] 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* (2018).

[51] Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. Help Me Write a Story: Evaluating LLMs' Ability to Generate Writing Feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25827–25847. doi:10.18653/v1/2025.acl-long.1254

[52] 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* (2025).

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

[54] Jose Ma Santiago, Richard Lance Parayno, Jordan Aiko Deja, and Briane Paul V. Samson. 2023. Rolling the Dice: Imagining Generative AI as a Dungeons & Dragons Storytelling Companion. arXiv:2304.01860 [cs] doi:10.48550/arXiv.2304.01860

[55] Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. CoE: A Clue of Emotion Framework for Emotion Recognition in Conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 23548–23563. doi:10.18653/v1/2025.acl-long.1148

[56] Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph. D. Dissertation.

[57] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19937–19947.

[58] 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)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 13419–13428. https://aclanthology.org/2024.lrec-main.1175

[59] Maria Teleki, Xiangjue Dong, Haoran Liu, and James Caverlee. 2025. Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models. In *ICWSM 2025*.

[60] 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17659–17681. doi:10.18653/v1/2024.emnlp-main.978

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

[62] Tuul Triyason. 2023. Exploring the Potential of ChatGPT as a Dungeon Master in Dungeons & Dragons tabletop game. In *Proceedings of the 13th International Conference on Advances in Information Technology* (Bangkok Thailand, 2023-12-06). ACM, 1–6. doi:10.1145/3628454.3628457

[63] 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*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 3665–3679. https://aclanthology.org/2025.findings-naacl.203/

[64] Alexander Vezhnevets, Fabricio Pardo, Junhyuk Oh, et al. 2023. Concordia: A Generative Framework for Scalable Agent-Based Simulations Using Large Language Models. *arXiv preprint arXiv:2312.03664* (2023).

[65] Kurt Vonnegut. [n. d.]. Kurt Vonnegut on the Shapes of Stories. YouTube Video. https://www.youtube.com/watch?v=GOGru_4z1Vc

[66] 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*. 1–16.

[67] Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024. A full-duplex speech dialogue scheme based on large language model. *Advances in Neural Information Processing Systems* 37 (2024), 13372–13403.

[68] Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025. 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)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1352–1391. https://aclanthology.org/2025.naacl-long.63/

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

[70] 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*. 1–4.

[71] Zizhen Wang, Jiangyu Pan, Duola Jin, Jingao Zhang, Jiacheng Cao, Chao Zhang, Zejian Li, Preben Hansen, Yijun Zhao, Shouqian Sun, and Xianyue Qiao. 2025. 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)*. Association for Computing Machinery, New York, NY, USA, Article 131, 21 pages. doi:10.1145/3706598.3713602

[72] 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* (2024).

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

[74] Zhaoyang Xia, Somdeb Sarkhel, Mehrab Tanjim, Stefano Petrangeli, Ishita Dasgupta, Yuxiao Chen, Jinxuan Xu, Di Liu, Saayan Mitra, and Dimitris N Metaxas. 2025. VISIAR: Empower MLLM for Visual Story Ideation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

[75] Kaige Xie and Mark Riedl. 2024. Creating Suspenseful Stories: Iterative Planning with Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2391–2407. doi:10.18653/v1/2024.eacl-long.147

[76] 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)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 16375–16398. doi:10.18653/v1/2025.acl-long.799

[77] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3378–3465. doi:10.18653/v1/2023.acl-long.190

[78] 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*. 3936–3951.

[79] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683* (2024).

[80] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. 2025. OASIS: Open Agent Social Interaction Simulations with One Million Agents. arXiv:2411.11581 [cs.CL] https://arxiv.org/abs/2411.11581

[81] 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*. 1–23.

[82] 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*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 1892–1907. https://aclanthology.org/2025.coling-main.129/

[83] Zaibin Zhang, Zhenfei Yin, and Jing Shao. 2024. GENSS: A GENERALIZED AND SCALABLE LLM-BASED AGENTS SOCIAL NETWORK SIMULATOR. In *NeurIPS 2024 Workshop on Open-World Agents*. https://openreview.net/forum?id=gS0hprhg72

[84] Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, and Wai Lam. 2025. The Essence of Contextual Understanding in Theory of Mind: A Study on Question Answering with Story Characters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 22612–22631. doi:10.18653/v1/2025.acl-long.1103

[85] Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Toronto, Canada, 2023). Association for Computational Linguistics, 11136–11155. doi:10.18653/v1/2023.acl-long.624

[86] Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. 2023. CALYPSO: LLMs as Dungeon Master's Assistants. 19, 1 (2023), 380–390. doi:10.1609/aiide.v19i1.27534

[87] Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. *Advances in Neural Information Processing Systems* 37 (2024), 100783–100815.

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 mainly focus on the conferences above, however, we also include additional works due to their unique and relevant contributions.

B Desired Features in Creative Support Tools

As a negative example, StoryNode [45] 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.

C The Role of Story in Persuasive Online Discourse

Previous social science research has established that people's beliefs and behaviors are impacted by storytelling, highlighting its value as a persuasive tool. In Nabhani et al. [32], they measure the presence of storytelling in `r/ChangeMyView`, a subreddit where users post their views on a subject, and sometimes change their views. `r/ChangeMyView` has a community norm wherein a person who has changed their view in response to a commenter re-comments with a Δ. Overall in Nabhani et al. [32], they found that storytelling is not highly predictive of persuasion in online discourse, in contrast to previous social science findings on in-person discourse.

D Related Tasks

In this section, we discuss tasks pertaining to stories, distinct from the story generation task.

D.1 Story Feedback

Rashkin et al. [51] investigate how effectively large language models can provide meaningful feedback on creative writing, in settings where the model can act as a writing coach rather than an automatic rewriter. They compare the quality of feedback (both positive comments and critical suggestions) produced by different LLMs for short stories, as well as their ability to identify obvious issues in the text.

To evaluate feedback quality, they consider four dimensions: (1) story-specificity, (2) usefulness for improving the story, (3) ability to detect the main issue, and (4) appropriate positive feedback when no obvious problems are present. They introduce StoryFeedback, a dataset of $\sim$84k pairs of English short stories and their model-generated feedback. It contains both original stories and their synthetically corrupted versions (via random sentence swaps, sentence deletion, or repeated English-German backtranslation). They benchmark eight publicly available LLMs (including Bloomz 7B/176 B, Gemma 9B/27B, Gemini 1.5 Flash/Pro and GPT 3.5/4 variants) under four prompting conditions: bullet point feedback,

bullet feedback guided by predefined issue categories, single-sentence feedback, and explicit error-spotting. For each prompt, they test zero-shot and two-shot settings.

The authors find that Gemma and Gemini models are most precise compared to others, at correctly judging when the story needs no changes. In general, this precision improves for larger models with two-shot prompts. Larger models also produce more diverse feedbacks overall, with Gemini and GPT 4 yielding the highest diversity. Human evaluation of feedbacks shows that while models perform strongly on specificity and reasonably on correctness, they are noticeably weaker at error-detection and giving relevant, issue-focused feedback. They often miss the main artificially introduced problem and sometimes incorrectly declare a story flawless. GPT 4 performs comparatively better on error-detection and relevance, and bulleted-list prompts (especially when aligned with issue categories) tend to yield high quality feedback. Models seem to able to detect issues introduced by backtranslation, but struggle with swapped or deleted sentences.

Although this study focuses on short stories with artificially generated errors and evaluates only a limited set of feedback dimensions, it provides a valuable first step toward a systematic evaluation framework for story-writing feedback. It highlights the need for LLMs that can identify the most salient writing issues, and calls for extending these evaluation frameworks to include long-form narratives with more complex writing problems.

D.2 Story QA

Zhou et al. introduce CHARTOM-QA [84], a benchmark that investigates how large language models (LLMs) handle Theory of Mind (ToM) when queried about characters whose mental states depend on long-term personal histories, rather than only on short, local situations. The authors argue that most existing ToM benchmarks rely on brief, synthetic stories that omit rich background information, while human ToM in real life draws heavily on extended biographies, social relations, and prior interactions. To study this gap, they build CHARTOM-QA: 1,035 ToM questions grounded in characters from 20 classic novels, targeting four standard ToM dimensions: belief, intention, emotion, and desire. The dataset is constructed via an AI-assisted pipeline around real reader highlights from public domain novels: user notes are filtered for explicit ToM content (manual checks suggest about 93 percent consistency with the stories), annotators extract the key phrase that encodes the ToM information, GPT-4o paraphrases this fragment into a complete description that annotators lightly edit, and GPT-4o then proposes candidate question and answer items which are machine filtered and human selected for a generative QA setup, with a multiple choice variant created by adding plausible distractors. Evaluation uses a grading-style protocol that derives discrete "bonus points" from reference answers and measures coverage, with penalties for defective content and for generative QA. It also measures accuracy for four-option multiple choice in both English and Chinese, under varying plot window lengths.

Experimental results show that GPT-4o is consistently the strongest model, but still covers only 35 to 50 percent of key bonus points and exhibits high penalty rates, indicating frequent inappropriate or defective reasoning. Longer plot windows slightly reduce penalties but do not reliably improve bonus point coverage or multiple choice accuracy, which suggests that current models do not effectively exploit extended narrative context for ToM. A human study on 150 multiple choice questions finds that participants who have read the novels outperform those who have not by about 21 percentage points and clearly surpass GPT-4o and stronger reasoning models such as o1 and DeepSeek-R1, while unfamiliar humans perform at roughly the same level as the models. Humans also benefit substantially from longer plot windows, while model performance remains largely flat. Further analysis shows that even reasoning focused models with explicit chain of thought often attempt, but fail, to incorporate character backstory and prior events, and that all models perform particularly poorly on "indirect" questions whose answers depend on global story knowledge. The authors conclude that CHARTOM-QA exposes a fundamental limitation of current LLMs: they struggle to integrate nuanced, long horizon contextual information about individuals when performing ToM reasoning, and the benchmark is intended to push work toward genuinely context sensitive, story level social understanding rather than pattern matching on short snippets.

In a related direction, CHARACTERCRITIQUE [71] 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 [7] 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.