StoryBox: Collaborative Multi-Agent Simulation for Hybrid Bottom-Up Long-Form Story Generation Using Large Language Models

Zehao Chen[1], **Rong Pan**[1]*, **Haoran Li**[2]*

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]School of Arts, Sun Yat-sen University, Guangzhou, China
chenzh593@mail2.sysu.edu.cn, panr@sysu.edu.cn, lihr83@sysu.edu.cn

Abstract

Human writers often begin their stories with an overarching mental scene, where they envision the interactions between characters and their environment. Inspired by this creative process, we propose a novel approach to long-form story generation, termed hybrid bottom-up long-form story generation, using multi-agent simulations. In our method, agents interact within a dynamic sandbox environment, where their behaviors and interactions with one another and the environment generate emergent events. These events form the foundation for the story, enabling organic character development and plot progression. Unlike traditional top-down approaches that impose rigid structures, our hybrid bottom-up approach allows for the natural unfolding of events, fostering more spontaneous and engaging storytelling. The system is capable of generating stories exceeding 10,000 words while maintaining coherence and consistency, addressing some of the key challenges faced by current story generation models. We achieve state-of-the-art performance across several metrics. This approach offers a scalable and innovative solution for creating dynamic, immersive long-form stories that evolve organically from agent-driven interactions.

Project — https://storyboxproject.github.io
Extended version — https://arxiv.org/abs/2510.11618

Introduction

The advent of large language models (LLMs) brings significant advancements to various fields, including multi-agent simulations (Li et al. 2024; Qian et al. 2024; Chen et al. 2024). These simulations offer a powerful tool for modeling complex interactions between virtual agents, providing a dynamic and context-rich environment for story generation. When humans write stories, they typically have an overarching mental picture of the story world. In our approach, multi-agent simulations are used to create this overarching scene, where agents (i.e., the characters in the story) interact based on predefined behaviors, triggering emergent events that form the foundation of the story. This allows for the automatic generation of diverse and engaging scenarios, which are crucial for building long-form stories.

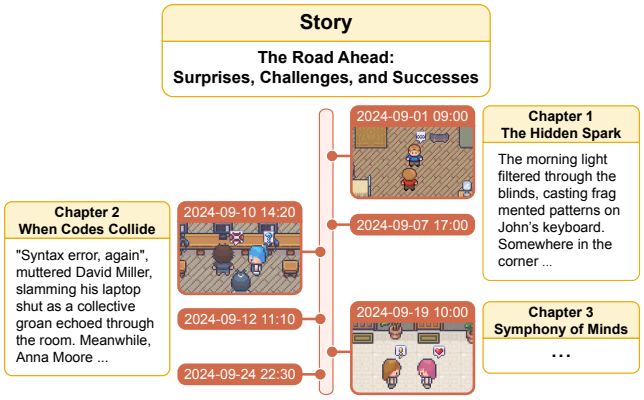

Figure 1: The timeline of the multi-agent sandbox simulation, where agent interactions with each other and their environment trigger emergent events that drive dynamic, hybrid bottom-up story generation.

As shown in Figure 1, the multi-agent sandbox simulation unfolds over time, with a series of events occurring within the sandbox. These events arise from interactions between agents, as well as between agents and their environment. In this process, agents act based on their predefined attributes, responding to and influencing the world around them. The interactions between agents and their environment give rise to a chain of events that continuously evolve, creating a dynamic storyline. The sandbox thus serves as a virtual space that mirrors the mental scene a human writer might envision when crafting a story. It is within this evolving sandbox that the foundations of the story are formed, from which a complete story can be generated based on the events that unfold.

In this setup, the sandbox serves a critical function: it is akin to the mental scene which a writer imagines, but with the difference that the interactions and outcomes are automatically generated through agent-based simulation rather than purely through human imagination. Each interaction, whether between characters or between characters and their surroundings, adds new layers to the story. By allowing agents to operate in this evolving and flexible sandbox, we ensure that the story's events are not preordained but emerge organically from the agents' behavior and environmental dynamics.

*Corresponding authors.

Traditional story generation often follows a top-down method (Zhou et al. 2023; Wang et al. 2023; Fan, Lewis, and Dauphin 2019; Goldfarb-Tarrant et al. 2020; Yang et al. 2022), where a story's structure is outlined first and then expanded. Although this framework provides a guide, it can limit natural character development and plot progression, resulting in stories that feel forced or predictable. In contrast, our hybrid bottom-up approach starts with agent-driven simulations that generate emergent events. As agents interact, these events unfold naturally, enabling more organic character growth and plot evolution, ultimately producing stories with greater depth, coherence, and spontaneity.

Our approach excels at generating longer, more detailed stories that stay consistent and cohesive. Using multi-agent simulations, it creates diverse, meaningful events that enrich plot and characters. Our main contributions are:

- A multi-agent simulation framework that generates dynamic and contextually rich story events through interactive agent behaviors.
- A hybrid bottom-up process that combines emergent interactions with guided storytelling to enable natural character development and plot progression.
- The ability to produce coherent and consistent long-form stories exceeding 10,000 words, effectively overcoming limitations of current story generation models.

Related Work

LLM-Based Multi-Agent Simulation

LLM-based multi-agent simulations have gained attention for their advanced language processing and decision-making, enabling nuanced agent interactions (Jinxin et al. 2023; Liu et al. 2023; Feng, Feng, and Qin 2023). These systems have been explored in various domains (Wang et al. 2024; Williams et al. 2023), particularly social simulation, which is most relevant to our research. Social simulation with LLM-based agents provides an effective way to model complex societal dynamics that are otherwise difficult or costly to study (Li, Yang, and Zhao 2023; Li et al. 2023). For example, S^3 (Gao et al. 2023) investigates the spread of information, emotions, and attitudes in social networks, while Generative Agents (Park et al. 2023) and AgentSims (Lin et al. 2023) model daily human interactions in virtual towns. Social Simulacra (Park et al. 2022) focuses on simulating online community regulation, and SocialAI School (Kovač et al. 2023) uses LLMs to simulate child development. These studies highlight the growing potential of LLM-based agents to offer valuable insights into societal behavior, community regulation, and social development, which provides a foundation for hybrid bottom-up long-form story generation.

Story Generation

Story generation has advanced with the rise of LLMs, which produce fluent stories but often struggle with coherence and consistency (Huot et al. 2024). To address these challenges, frameworks such as outlining followed by expansion into full stories have been proposed (Zhou et al. 2023; Wang et al. 2023). However, long-form story generation remains challenging. Recent studies have explored multi-agent systems using LLMs (Wang, Zhou, and Ledo 2024; Nasir, James, and Togelius 2024), such as Agents' Room (Huot et al. 2024), which focuses on generating stories of 1,000-2,000 words using Planning and Writing Agents coordinated by an Orchestrator, and IBSEN (Han et al. 2024), which employs Director-Actor agent collaboration for scriptwriting. While top-down approaches, such as outlining, provide structure, they limit natural storylines and character interactions. In contrast, our hybrid bottom-up approach generates dynamic stories from simulations, producing richer and more coherent stories of over 10,000 words, with clear advantages over traditional methods.

Methodology

This section introduces our long-form story generation system (Figure 2). It has two parts: a sandbox simulating multi-agent interactions to create events, and a Storyteller Agent that turns these events into coherent, engaging stories.

Multi-Agent Simulation

The multi-agent simulation serves as the foundation for generating long-form stories. Both the process of long-form story generation and standalone multi-agent simulation rely on well-defined character settings. Inspired by the character modeling approach of Generative Agents (Park et al. 2023), we adopt and modify it to better suit the specific requirements of story generation. This modification ensures that the characters exhibit coherent behavior and contribute meaningfully to the overall story structure.

Core Attributes As illustrated in Figure 2 "Persona Scratch Information", we first define the initial settings for each character. These initial settings are critical for the agents to engage in realistic interactions and generate compelling events. Each character is defined by a set of core attributes, which include Name, Age, Innate, Learned, Currently, and Lifestyle. These attributes describe both static and dynamic aspects of the character, providing a foundation for their behaviors and decisions. To model the character's daily actions, we introduce an additional attribute, Daily Plan Requirements. This attribute contains specific tasks or routines that the character plans to accomplish, such as conducting research, reading papers, etc. Importantly, this attribute is dynamic: at the start of a new day, a fresh set of daily plans is generated. Based on these daily plan requirements, a detailed schedule is created, with tasks assigned to specific hours of the day. This enables the character to follow a structured routine, while leaving room for flexibility and variation.

Abnormal Behavior Attribute To add depth and excitement to the generated stories, it is crucial to incorporate elements that break from routine. A monotonous, predictable character would lead to dull and uneventful stories. Therefore, we introduce an Abnormal Behavior attribute for each character. This attribute determines whether the character is likely to engage in actions that deviate from their usual behavior, such as abandoning their daily plan to pursue other

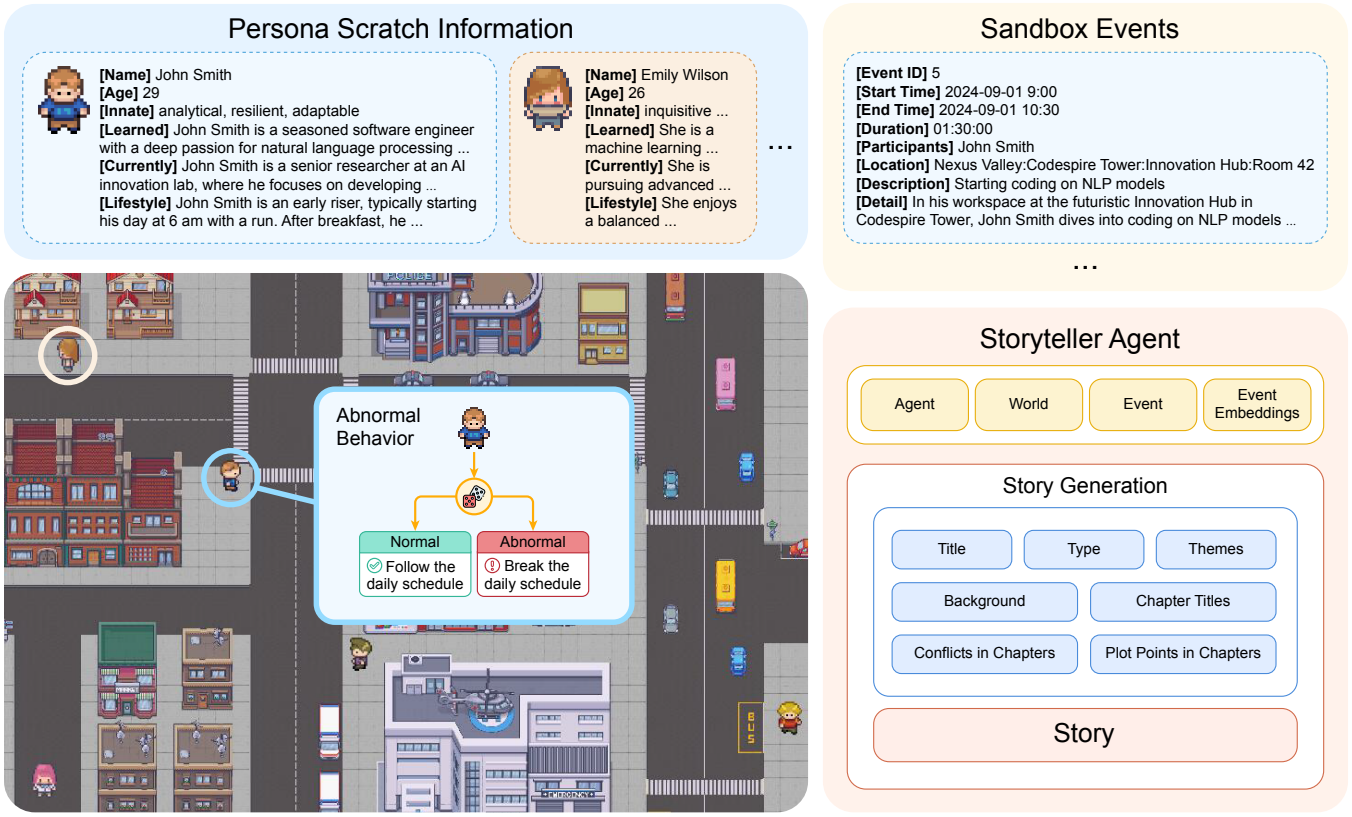

Figure 2: Overview of the system framework for long-form story generation, including the Persona Scratch Information for defining character settings, the sandbox where agent interactions generate events, and the Storyteller Agent that uses these events to craft a complete story.

activities or engaging in conflicts with other characters. The probability of a character exhibiting abnormal behavior is controlled by a hyperparameter, the Abnormal Factor. A higher value for this factor increases the likelihood that the character will break from their routine and introduce more dynamic and unpredictable actions.

Character Behavior Types Character behavior is categorized into three main types: move, chat, and none. The move behavior indicates that the character physically moves to a different location and performs specific actions. The chat behavior represents the character engaging in conversations with other characters. The none behavior means the character does nothing, allowing for moments of reflection, rest, or inaction, which contribute to pacing and tension within the simulation. These behaviors are selected based on the character's current attributes and the context of the simulation.

Together, these attributes define a character's personality, daily routines, and potential for unpredictability, creating a rich and dynamic environment for simulation. The agents interact with each other based on these characteristics, generating events that reflect plausible interpersonal dynamics and responses to environmental changes. This simulation framework serves as the raw material for long-form story generation, with the emergent events providing the foundation for compelling stories.

Event Recording In the sandbox, every action performed by an agent is considered an event, as illustrated in Figure 2 under "Sandbox Events". Each event is assigned a unique ID within the sandbox, along with a defined start time and end time, allowing us to calculate its duration. Events can involve multiple participants; for example, a conversation event will always have at least two participants. Additionally, each event must be recorded with its location, a brief description of the event (referred to as description), and a more detailed account (referred to as detail) to capture the full context.

The description attribute provides a concise summary of the event, such as "Starting coding on NLP models", which briefly explains what the event is about. However, this brief description is often insufficient for generating a rich and engaging story, as it lacks important contextual details. To address this, we introduce the detail attribute, which expands upon the basic description and includes additional information necessary for story generation. The detail attribute incorporates factors such as the environment, the time of day, the status of the characters involved, and the location of the event. Furthermore, it captures character-specific elements, including their actions, emotional states, and possible motivations during the event.

By providing these details, we ensure that the events

within the sandbox are not just isolated actions, but are rich, contextually grounded occurrences that can later be woven into the story. This level of granularity allows the Storyteller Agent to craft stories that are nuanced and immersive, as the events are not only linked to character behaviors but are also informed by the surrounding context and dynamics.

Through this multi-agent simulation, we create a dynamic sandbox environment where interactions between characters generate a wide variety of events. These events, rich in context and detail, form the core elements for story generation. The Storyteller Agent then processes these events, transforming them into intricate, character-driven long-form stories that maintain both depth and coherence in the story.

Environment Modeling The environment plays a critical role in the sandbox. In Generative Agents, the environment is modeled using a tile-based system, where tiles represent the basic environmental units commonly found in RPG games. However, this approach has inherent limitations, particularly the need for precise coordinates, which can make modeling more rigid and less flexible. To overcome these limitations, we adopt a more general tree-like structure for environment modeling, which does not rely on specific coordinates but instead uses relative distances between objects and areas. For a detailed introduction to environment modeling, please refer to the supplementary material in the extended version.

Unlike the fixed, grid-based environment of Generative Agents, which confines agents to a limited simulation space, our model enables a virtually limitless environment. This means that the sandbox can be as large and complex as needed, without the constraint of pre-defined spatial boundaries. Characters can explore a more expansive and immersive world, interacting with a wider range of objects and areas, making the simulation more dynamic and fluid.

Hybrid Bottom-Up Long-Form Story Generation

Once the events generated through the multi-agent sandbox simulation are available, the next step is to use the Storyteller Agent to generate the long-form story. As illustrated in Figure 3, this process follows a structured workflow that incorporates several key stages.

Event Summarization The sandbox events are arranged chronologically, but the sheer number of events can be overwhelming. Directly feeding all these events into a LLM is not feasible due to the limitations of the model's context window. Therefore, summarization is necessary. First, we summarize the events by character, recording what each character did on each day. This summary is also arranged chronologically. Once we have these character-based daily summaries, we introduce a dynamic windowing mechanism that groups events into smaller chunks for summarization. The size of the dynamic window is automatically determined by the LLM, allowing for adaptability to the complexity and density of the events. The summaries within each window are abstract and condensed, reducing the volume of data while maintaining chronological order. This layered and dynamic approach ensures that the event data remains manageable for the LLM while preserving key information.

Story Information Generation In the story generation process, we first generate the story type (e.g., adventure, mystery, romance) rather than beginning with the title. The story type plays a crucial role in shaping the story, as it provides a foundational framework that guides the subsequent development of the plot, characters, and themes. By determining the story type at the outset, we establish a clear direction for the story, ensuring coherence and consistency throughout the story's progression.

After determining the story type, we generate the story title beginning with the initial event summary. The title is iteratively refined with each new event summary, allowing the model to update the title dynamically based on evolving context. This process not only produces a cohesive and relevant title but also acts as a filter, highlighting key elements and capturing the story's essence. The language model decides whether to retain or modify the title depending on the current event summaries.

Once the title is finalized, a secondary selection process identifies the most suitable title. This can involve using a language model for evaluation, manual review, or training a dedicated model. Due to the lack of a specialized dataset, we opt for a fully automated LLM-based approach that ranks titles based on relevance, creativity, and coherence with the event summaries.

Following title generation, other key story information is produced, including background, main themes, chapter titles, conflicts per chapter, and major plot points. These elements are informed by the story type, title, and prior event summaries, ensuring a well-structured and thematically consistent story.

If manual inputs are provided, the Storyteller Agent skips automatic generation of these elements. Additionally, hyperparameters like the number of themes, chapters, conflicts per chapter, and plot points per chapter allow customization of the story's scope and depth.

Story Generation Once the story information is ready, the actual story generation process begins. A key component in this stage is the information retrieval module, which receives two types of input data: story-related information (such as title, themes, and chapter details) and sandbox data (including character information, environment details, and the events themselves). To further enhance event matching for information retrieval, we convert events into event embeddings using an embedding model. This allows events to be retrieved both through keyword-based search and dense vector-based retrieval.

It is important to highlight that the process of information retrieval and its application in story generation is inherently bottom-up in nature. This is because the events from the sandbox, which play a crucial role in the generation process, are drawn upon, distinguishing this approach from traditional top-down story generation methods. Furthermore, this information retrieval process also acts as a dynamic filtering mechanism, automatically selecting meaningful events that align with the story's progression. By continuously refining the event selection based on the evolving story, the system ensures that only the most relevant and engaging content is

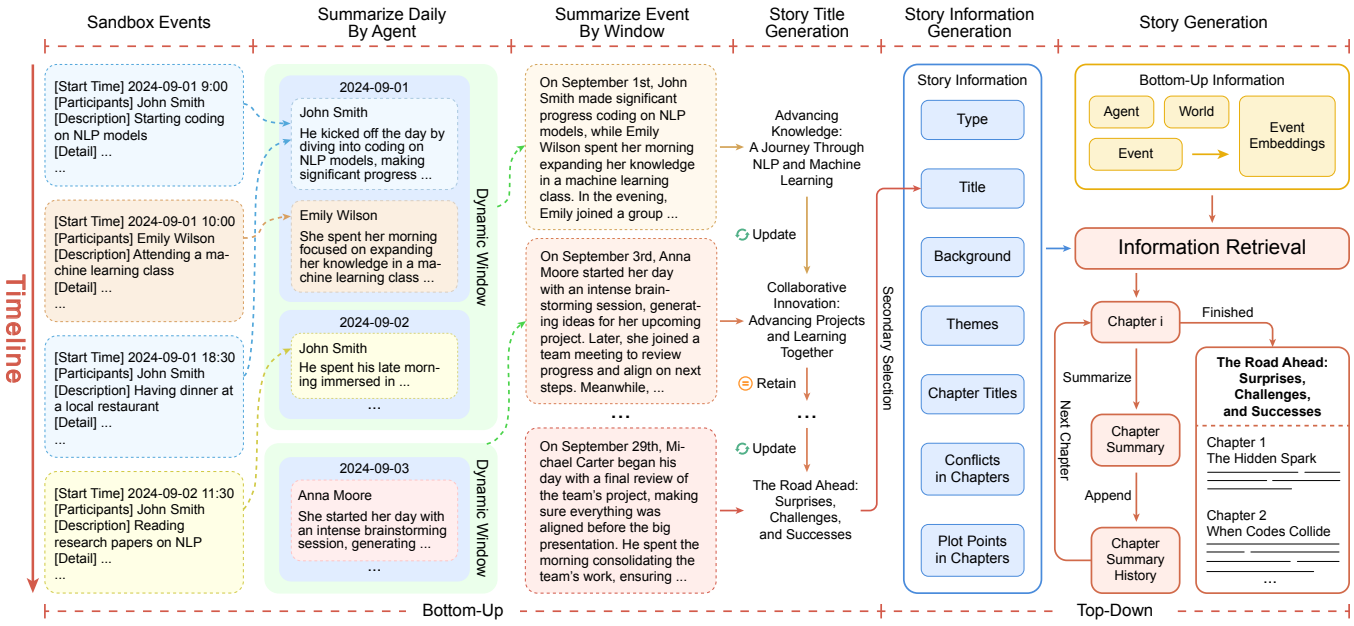

Figure 3: Overview of the Storyteller Agent workflow for generating long-form story using a hybrid bottom-up approach, from sandbox events to iterative story generation.

used to shape the story.

The story generation process is driven by an iterative loop. During the generation of the first chapter, the system retrieves relevant information, such as the story title, themes, and chapter details, to serve as the foundation for crafting the content. It's important to note that a chapter may not be fully generated in a single pass; instead, it undergoes several iterations before it is completed. Once a chapter is finished, it is summarized, and the summary is added to the chapter summary history, which serves as input for generating the next chapter.

Each subsequent chapter's generation includes not only the information retrieved for the current chapter but also the summaries of the previous chapters. This iterative process continues, chapter by chapter, until the entire story is completed. This hybrid bottom-up method allows for the generation of long, coherent stories, where each chapter builds on the events and summaries that have come before, ensuring continuity and story flow.

Experiments

In this section, we present experiments to validate the effectiveness of StoryBox, including evaluation metrics, baseline comparisons, and result analysis.

Experimental Settings

Dataset Following the setup described in DOC (Yang et al. 2023), we use premises, settings, and characters as inputs. However, the DOC dataset contains relatively homogeneous story types, so we construct a new dataset consisting of 20 story-related settings without reference answers. Our dataset features a broader variety of story types, including genres such as science fiction. For detailed information about our dataset, please refer to the supplementary material in the extended version.

Evaluation Metrics Evaluating generated stories is inherently challenging due to the subjective, multi-dimensional, and context-dependent nature of narrative quality (Guan and Huang 2020; Chhun et al. 2022; Yang and Jin 2024). Inspired by recent advances such as Agents' Room (Huot et al. 2024), we adopt a hybrid evaluation strategy that combines both human and automated methods to provide a comprehensive assessment of story quality. Building upon this foundation, we expand the evaluation criteria to better suit the story generation setting.

(1) Human Evaluation We conduct structured pairwise comparisons across six criteria: Plot, Creativity, Character Development, Language Use, Conflict Quality, and Overall. These dimensions cover key aspects of long-form storytelling, including narrative coherence, originality, emotional depth, and stylistic quality. Annotators with backgrounds in literature, storytelling, and related fields compare story pairs from the same setting and provide ratings for each dimension as well as an overall preference. To reduce fatigue and bias, each annotator evaluates a limited number of pairs per session. We apply a Latin Square design to balance the assignment of story settings and randomize story order. For each setting, we collect $N \times (N-1)/2$ comparisons, where N is the number of methods.

(2) Automatic Evaluation We also use a large language model to perform evaluations based on the same criteria as human judgments. The model is guided by structured prompts aligned with human instructions and compares story pairs across key dimensions. This approach offers a cost-effective approximation of human evaluation and

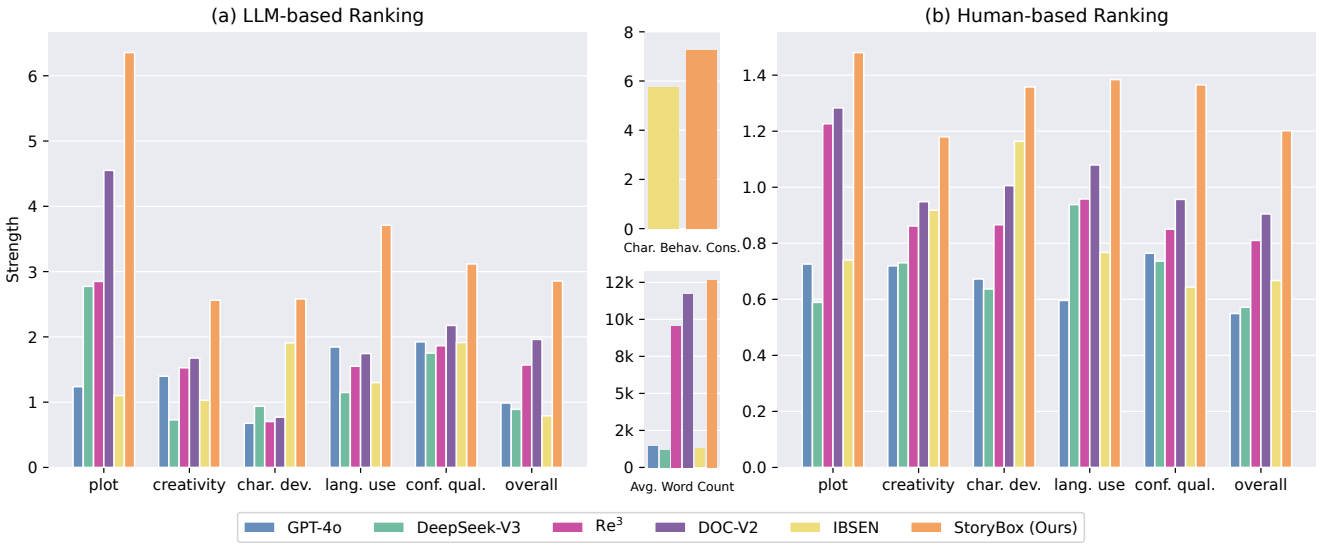

Figure 4: Comparative performance of different methods across multi evaluation dimensions: Plot, Creativity, Character Development, Language Use, Conflict Quality, and Overall. Subfigure (a) presents rankings based on LLM-based evaluation, while subfigure (b) shows rankings from human evaluation. We also include the sandbox-specific metric Character Behavior Consistency, along with the Average Word Count for each method.

allows us to assess its alignment with human ratings. In addition, we include a sandbox-specific metric, **Character Behavior Consistency**, which evaluates how well characters act in line with their defined personas. We also report **Average Word Count** to reflect story length.

For full metric definitions, evaluation procedures, and interface details, please refer to the supplementary material in the extended version.

Baselines We select several representative and high-performing methods as our baselines. We categorize the baseline methods into three types: (1) **Vanilla LLMs**: methods that generate long-form stories directly using large language models, such as GPT-4o (Hurst et al. 2024) and DeepSeek-V3 (Liu et al. 2024); (2) **Structured Frameworks**: methods that utilize structured frameworks for long-form story generation based on LLMs, such as Re^3 (Yang et al. 2022) and DOC-V2 (Yang et al. 2023); and (3) **Multi-Agent Simulations**: methods that generate through multi-agent simulations, such as IBSEN (Han et al. 2024).

Implementation Details The process of implementing the system, including the story-related settings, sandbox initialization, and the setup of characters and environment, is described in the supplementary material of the extended version. The specific prompts used in this paper, along with additional implementation details, are also provided therein.

Automatic Evaluation

We conduct an automatic evaluation of several methods, with results shown in Figure 4(a). StoryBox consistently achieves the best performance across all major metrics.

For the **Plot** metric, it maintains strong coherence, while IBSEN performs less well, likely due to its lack of narrative-specific design. Most models show similar results in **Creativity**, except DeepSeek-V3, which performs noticeably worse, suggesting limited stylistic diversity. In **Character Development**, StoryBox ranks highest, and IBSEN follows, indicating that sandbox-based simulations are effective in modeling character dynamics. StoryBox also leads in **Language Use**, thanks to the Storyteller Agent's integration of environmental and contextual cues, which enriches the narrative with vivid descriptions. For **Conflict Quality**, it performs best, reflecting its ability to construct structured and engaging tension. The **Overall** score confirms StoryBox's strong and balanced performance across dimensions. In the sandbox-specific **Character Behavior Consistency** metric, StoryBox slightly outperforms IBSEN, suggesting that simulation contributes to coherent agent behavior. Finally, in the **Average Word Count** metric, which reflects long-form generation ability, StoryBox produces stories averaging around 12,000 words. Other long-form methods typically generate about 10,000 words, while models without such adaptation produce much shorter texts, averaging only around 1,000 words, revealing a clear limitation in generating extended story.

Additional results, including case studies, are provided in the supplementary material of the extended version.

Human Evaluation

Given the inherently subjective nature of story generation, human evaluation is essential for a comprehensive assessment. To capture diverse reader perspectives, we recruited 78 participants from a variety of academic backgrounds, including literature, computer science, and other disciplines. Each participant was asked to compare pairs of stories generated from the same setting, based on the same set of di-

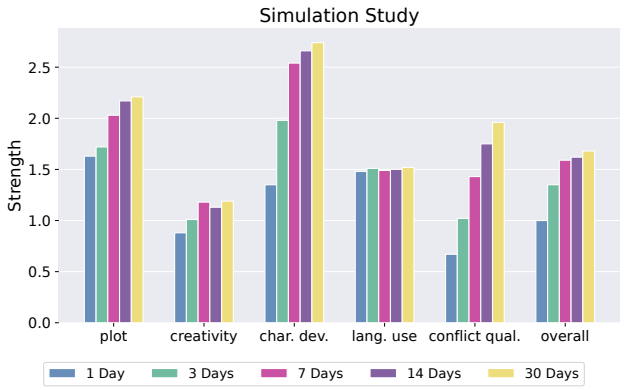

Figure 5: Effect of simulation duration on story generation performance.

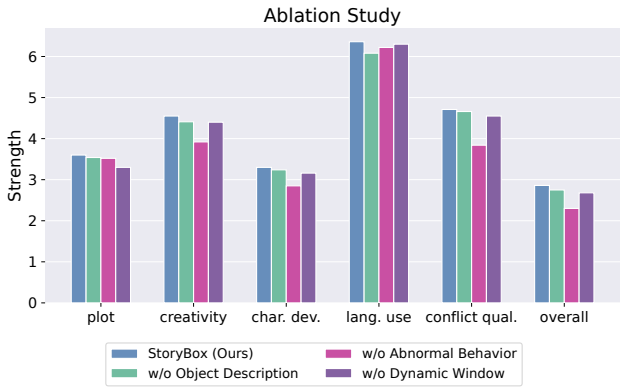

Figure 6: Performance comparison of StoryBox without different components.

mensions used in the automatic evaluation.

The human evaluation results are presented in Figure 4(b). Overall, StoryBox consistently achieves the highest scores across all dimensions, indicating strong performance in both structural and stylistic aspects of storytelling. This further confirms the effectiveness of the system across a range of reader expectations. Compared to the automatic evaluation results in Figure 4(a), the trends are largely aligned, suggesting that automatic metrics provide a reasonably accurate reflection of human judgments. Nonetheless, some discrepancies are observed, as human evaluators show a stronger preference for IBSEN over GPT-4o and DeepSeek-V3. This highlights the possibility that simulation-based narratives may offer human-relatable qualities that are not fully captured by automated metrics.

Simulation Duration Study

We investigate how varying simulation durations (1, 3, 7, 14, and 30 days) affect story quality, while keeping story length fixed at around 12,000 words. As shown in Figure 5, not all metrics benefit equally from longer simulations.

Plot, **Creativity**, and **Language Use** show minimal variation across durations, likely due to the strong guidance of the Storyteller Agent and the language model's inherent capabilities. In contrast, **Character Development** and **Conflict Quality** improve with longer simulations, as extended interactions allow for deeper character arcs and more complex tensions. The **Overall** metric improves significantly from 1 to 7 days, but shows diminishing returns beyond that. Notably, token usage roughly doubles with each duration increase, while quality gains plateau. Excessive events in longer simulations may even burden the model's ability to synthesize information effectively.

Overall, a 7-day simulation strikes a practical balance between quality and efficiency, making it a suitable default under current model constraints.

Ablation Study

We conduct an ablation study to assess the impact of three components: (1) object descriptions, (2) random abnormal behaviors, and the (3) dynamic context window. As shown in Figure 6, removing any component leads to a performance drop, though to varying degrees.

Excluding **object descriptions** harms **Language Use**, as environmental details help the Storyteller Agent produce richer and more grounded narratives. Removing **random abnormal behaviors** causes the sharpest decline, particularly in **Creativity**, **Character Development**, and **Conflict Quality**, highlighting the importance of unpredictability in driving dynamic stories. Disabling the **dynamic context window** reduces **Plot** quality, as it limits the model's ability to maintain coherence across longer narrative arcs. Overall, each component plays a distinct role, and their combination is key to StoryBox's effectiveness.

Limitations

While our multi-agent sandbox-based approach shows strong potential, it still has limitations. First, the simulation currently proceeds sequentially, with each character acting one at a time, which slows down the overall process. Parallelizing the simulation could improve efficiency, but managing interdependent character behaviors remains a challenge. For instance, one character's actions may directly affect others, and naive parallelization could introduce inconsistencies. Second, evaluating story quality is still a difficult problem. Although we use both human and automatic evaluations, human judgment is costly and time-consuming. More accurate and scalable automated metrics are needed to better approximate human preferences and enable broader application. These issues represent important directions for future work.

Conclusion

In this paper, we introduced StoryBox, a novel approach for long-form story generation using multi-agent simulations. Our experiments show that StoryBox outperforms existing methods on many metrics. Despite challenges in evaluating story quality, both automatic and human evaluations confirm its effectiveness in generating engaging and coherent stories.

Acknowledgements

The authors are supported by the National Key Research and Development Program of China (Grant No. 2020YFA0712500). The authors would also like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

References

Chen, J.; Hu, X.; Liu, S.; Huang, S.; Tu, W.-W.; He, Z.; and Wen, L. 2024. LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13055–13077. Bangkok, Thailand: Association for Computational Linguistics.

Chhun, C.; Colombo, P.; Suchanek, F. M.; and Clavel, C. 2022. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 5794–5836. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Fan, A.; Lewis, M.; and Dauphin, Y. 2019. Strategies for Structuring Story Generation. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2650–2660. Florence, Italy: Association for Computational Linguistics.

Feng, X.; Feng, X.; and Qin, B. 2023. The Role of Summarization in Generative Agents: A Preliminary Perspective. *arXiv preprint arXiv:2305.01253*.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Goldfarb-Tarrant, S.; Chakrabarty, T.; Weischedel, R.; and Peng, N. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4319–4338. Online: Association for Computational Linguistics.

Guan, J.; and Huang, M. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9157–9166. Online: Association for Computational Linguistics.

Han, S.; Chen, L.; Lin, L.-M.; Xu, Z.; and Yu, K. 2024. IBSEN: Director-Actor Agent Collaboration for Controllable and Interactive Drama Script Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1607–1619. Bangkok, Thailand: Association for Computational Linguistics.

Huot, F.; Amplayo, R. K.; Palomaki, J.; Jakobovits, A. S.; Clark, E.; and Lapata, M. 2024. Agents' Room: Narrative Generation through Multi-step Collaboration. *arXiv preprint arXiv:2410.02603*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jinxin, S.; Jiabao, Z.; Yilei, W.; Xingjiao, W.; Jiawen, L.; and Liang, H. 2023. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503*.

Kovač, G.; Portelas, R.; Dominey, P. F.; and Oudeyer, P.-Y. 2023. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871*.

Li, C.; Su, X.; Han, H.; Xue, C.; Zheng, C.; and Fan, C. 2023. Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

Li, N.; Gao, C.; Li, M.; Li, Y.; and Liao, Q. 2024. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15523–15536. Bangkok, Thailand: Association for Computational Linguistics.

Li, S.; Yang, J.; and Zhao, K. 2023. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *arXiv preprint arXiv:2307.10337*.

Lin, J.; Zhao, H.; Zhang, A.; Wu, Y.; Ping, H.; and Chen, Q. 2023. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Zhou, D.; Dai, A. M.; Yang, D.; and Vosoughi, S. 2023. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Nasir, M. U.; James, S.; and Togelius, J. 2024. Word2World: Generating Stories and Worlds through Large Language Models. *arXiv preprint arXiv:2405.06686*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Qian, C.; Dang, Y.; Li, J.; Liu, W.; Xie, Z.; Wang, Y.; Chen, W.; Yang, C.; Cong, X.; Che, X.; Liu, Z.; and Sun, M. 2024. Experiential Co-Learning of Software-Developing Agents. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5628–5640. Bangkok, Thailand: Association for Computational Linguistics.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wang, Y.; Yang, K.; Liu, X.; and Klein, D. 2023. Improving Pacing in Long-Form Story Planning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10788–10845. Singapore: Association for Computational Linguistics.

Wang, Y.; Zhou, Q.; and Ledo, D. 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*, FDG '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400709555.

Williams, R.; Hosseinichimeh, N.; Majumdar, A.; and Ghaffarzadegan, N. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Yang, D.; and Jin, Q. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation. *arXiv preprint arXiv:2408.14622*.

Yang, K.; Klein, D.; Peng, N.; and Tian, Y. 2023. DOC: Improving Long Story Coherence With Detailed Outline Control. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3378–3465. Toronto, Canada: Association for Computational Linguistics.

Yang, K.; Tian, Y.; Peng, N.; and Klein, D. 2022. Re3: Generating Longer Stories With Recursive Reprompting and Revision. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4393–4479. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhou, W.; Jiang, Y. E.; Cui, P.; Wang, T.; Xiao, Z.; Hou, Y.; Cotterell, R.; and Sachan, M. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*.