BOOKWORLD: From Novels to Interactive Agent Societies for Creative Story Generation

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have enabled social simulation through multi-agent systems. Prior efforts focus on agent societies created from scratch, assigning agents with newly defined personas. However, simulating established fictional worlds and characters remain largely underexplored, despite its significant practical value. In this paper, we introduce BOOKWORLD, a comprehensive system for constructing and simulating book-based multi-agent societies. BOOKWORLD's design covers comprehensive real-world intricacies, including diverse and dynamic characters, fictional worldviews, geographical constraints and changes, *e.t.c.* BOOKWORLD enables diverse applications including story generation, interactive games and social simulation, offering novel ways to extend and explore beloved fictional works. Through extensive experiments, we demonstrate that BOOKWORLD generates creative, high-quality stories while maintaining fidelity to the source books, surpassing previous methods with a win rate of 75.36%. The code of this paper can be found at: https://anonymous.4open.science/r/BookWorld-1DC5/

1 Introduction

> *"Now we know that we are living our lives in a book ."*
> *"If what you say is true, I'm going to run away from the book and go my own way."*
> – *Sophie's World*, by Jostein Gaarder.

Characters in fictional works, once established, should take on a life of their own, beyond the manipulation of their authors. Bringing fictional characters and worlds into life has long captivated the imagination of fiction enthusiasts. Leveraging recent advances in large language models (LLMs), we can now craft interactive artificial society for characters in books through multi-agent systems (Park et al., 2023), which simulate social interactions of humans, facilitating diverse applications such as interactive games (Wu et al., 2024), social simulations (Zhou et al., 2023) and story creation (Han et al., 2024; Chen et al., 2024).

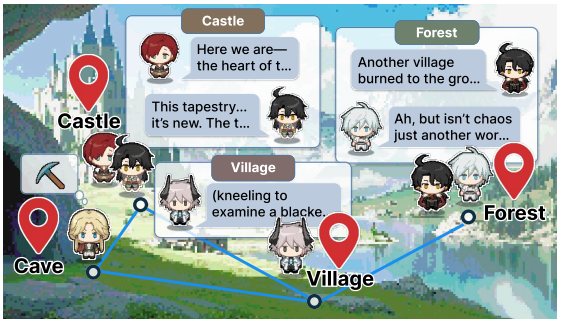

Figure 1: A preview of BOOKWORLD. Characters located at the same place may interact with each other based on their goals and other factors. The map is represented in a discrete form, and characters can spend a specified number of turns to move to adjacent locations.

Previous efforts primarily focus on creating agent societies from scratch, where agent personas are newly defined by brief descriptions or demographic traits, including: *1)* social simulations that study agents' social behaviors, ranging from specified scenarios such as debates (Chan et al., 2023) and strategic games (Wang et al., 2023a) to open-ended society simulation (Park et al., 2023; Dai et al., 2024); and *2)* task-oriented multi-agent systems, such as collaborative coding (Huang et al., 2023) and story generation (Han et al., 2024), where agents are specialized for different subtasks. However, agent societies simulating established fictional worlds remain underexplored.

In this work, we propose BOOKWORLD, a comprehensive system for book-based multi-agent societies, which simulates story progression in fictional worlds and facilitates story creation. BOOKWORLD extracts character data and background knowledge from source books, and constructs a multi-agent system using these data, comprising

role agents for the characters and a world agent for simulation control. The simulation progresses through individual scenes, where role agents of involved characters engage in various interactions such as working, communicating and trading. They continuously update their memories, status, and goals. The world agent orchestrates the simulation by managing system workflow, maintaining global status, providing environmental feedback, *e.t.c*. When the simulation ends, its histories weave together the stories, which are then polished by LLMs into cohesive, novel-style narratives. Our system also supports human intervention, *i.e.*, controlling the simulation via user-specified plots or scripts.

In addition, we systematically collect worldview data from books. As fictional works often contain various fantastical background elements, we extract comprehensive worldview data to enrich BOOKWORLD, including social norms, cultural contexts, and terminology explanations. These data enable agents in BOOKWORLD to act appropriately under corresponding worldviews.

The main contributions of this work are as follows:

- To the best of our knowledge, this work presents the first study of book-based agent societies for fictional world simulation. Such simulation allows character-driven storytelling, thus facilitating creative story generation.
- We introduce BOOKWORLD, a comprehensive framework for constructing and simulating book-based agent societies, covering systematic methodologies for data preparation, simulation, and rephrasing.
- We evaluate BOOKWORLD through comprehensive quantitative and qualitative analyses. The results demonstrate that BOOKWORLD generates high-quality narratives while maintaining fidelity to the source materials, outperforming previous methods in 75.36% cases.

2 Related Work

2.1 Multi-agent System

World Simulation Multi-agent systems for world simulation are developed to explore social dynamics in real-world applications. Such simulations could test social science theories within a small scope (Chuang et al., 2024) or populate virtual spaces and communities with large-scale realistic social phenomena.

Park et al. (2023) develops generative agents within a Sims-like interactive sandbox environment, enabling users to engage with a small community of 25 agents through natural language interactions. Project Sid (AL et al., 2024) deploys nearly a thousand agents with distinct personalities in Minecraft, establishing an "AI civilization" with complex human social activities such as trading and elections. Yang et al. (2024) develops OASIS, a large-scale social media simulation framework capable of orchestrating interactions among millions of agents.

Story Generation Some works have made the efforts to leverage multi-agent systems for story generation (Chen et al., 2024; Han et al., 2024). In these works, stories are generated by writing and expanding pre-defined outlines through constructing author (or director), editor, and actor agents to collaboratively solve the story generation task.

While these works demonstrate significant potential in world simulations and story generation, there remains limited exploration of book-based virtual environment construction and its application to story creation.

2.2 Automated Story Writing

Early research methods emphasize algorithm planning based on character traits and social constraints (Meehan, 1977; Lebowitz, 1984). With the rise of neural networks, research has gradually shifted towards data-driven machine learning methods (Yao et al., 2019; Goldfarb-Tarrant et al., 2020).

The advent of large language models has significantly transformed the landscape of automated story writing (Yuan et al., 2022). Most contemporary approaches employ a top-down methodology, generating narratives from the perspective of the writer or director, first creating high-level story outlines and then incrementally developing detailed content to enhance the overall quality of the narrative generation. While some approaches rely on a single large language model (Mirowski et al., 2023), others leverage multi-agent architectures to generate stories (Han et al., 2024; Chen et al., 2024).

However, experimental results show that these approaches still lag behind human professional writers in intrinsic creativity and textual complexity (Tian et al., 2024), often generating stories that

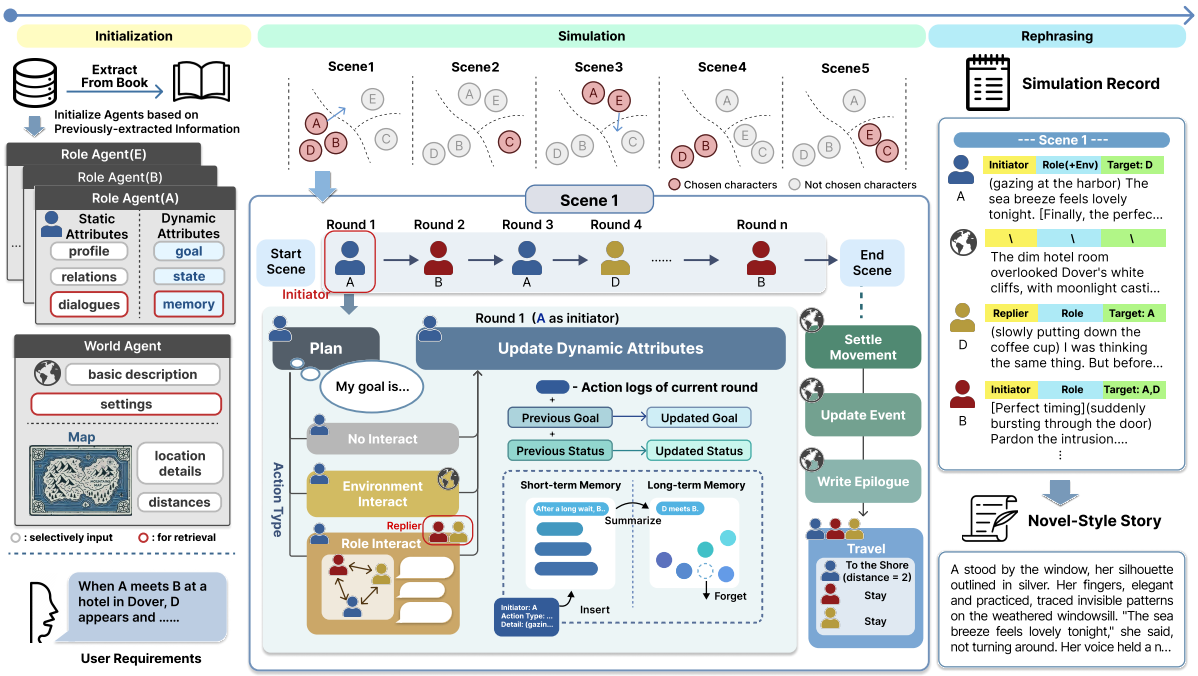

Figure 2: The complete pipeline of BOOKWORLD simulation. The framework starts from initializing the agents, then the agents interact across scenes.

lack suspense and tension and tend to produce homogenized, creatively lacking content (Gómez-Rodríguez and Williams, 2023; Ismayilzada et al., 2024).

3 BOOKWORLD

In this section, we elaborate on the design of BOOKWORLD. The primary motive of BOOKWORLD is to build a believable interactive environment for simulation based on available information, which includes two key objectives:

- Enhance the immersion of the constructed environment;
- Strengthen character autonomy within the environment while maintaining fidelity to their personality and experiences.

To achieve these objectives, we concentrate on data extraction and the design of agent interactions. Our data extraction process transforms original text into structured data, which forms the foundation for building our agents. In addition, we innovatively develop a specialized method to collect worldview details, ensuring adherence to global norms. Characters in the virtual world are granted substantial freedom. They can explore their environment, interact with other characters, and respond to various stimuli. A dynamic memory mechanism allows them to reflect on their experiences throughout simulations, all while preserving character fidelity and consistency.

3.1 Overview

The overall pipeline of simulation is illustrated in Figure 2.

Data Preparation Before the simulation starts, we extract character and worldview data from source materials, used for book-to-system construction. The extraction method is detailed in §3.4.

Simulation The simulation begins by initializing role agents and the world agent, loading character profiles, the map, the worldview and other necessary information. Each character establishes long-term motivations representing fundamental aspirations, such as defending their country. Details about the agents are introduced in §3.2.

Upon completing initialization, the simulation phase commences. Referring to dramatic theory (McKee, 1999), we define the minimal narrative unit as a **scene** to preserve narrative integrity and coherence. A scene is a bounded segment, analogous to a chapter in novels. The simulation continues for a user-specified number of scenes. Scenes retain modular independence, and they collectively compose an integral, cohesive narrative.

Within each scene, a selected group of role agents take turns playing as the initiator, making interactions with others or the environment based on their goals. Detailed scene arrangements are described in §3.3.

When the current scene ends, characters can travel to different locations. Those who are already traveling will reach their destination after a specified number of scenes have passed. The world agent tracks the progress of the current scene through updating events, and then selects characters for the next scene based on historical records.

Rephrasing After the simulation ends, we collect the simulation records and apply LLMs to rephrase the records into the final, novel-style story.

3.2 Architecture

Our approach empowers characters to act within a spatially and culturally grounded virtual world, enabling stories to emerge naturally from their interactions. To achieve this, we introduce two core components: **role agents** and **world agent**. These components are designed to work together, bridging individual character motivations with global environmental constraints, thereby ensuring both narrative creativity and logical coherence.

3.2.1 Role Agents

Role agents are the core of BOOKWORLD. They are able to make actions based on their intrinsic traits and exhibit complex social behaviors, forming individual motivations and memories during the simulation.

Attributes Role attributes contain fundamental traits essential for agent construction. These attributes are categorized into static and dynamic types:

- **Static Attributes** Static attributes include inherent characteristics such as gender, age, appearance, and personality. These attributes, specified in character profiles, remain constant throughout the simulation to maintain character consistency.
- **Dynamic Attributes** Dynamic attributes mainly include goals, states and memories. Initialized at the beginning, these attributes evolve with story progression, enabling dynamic character development.

> Right, Mr. Lorry. (Nods firmly) I'll see to it at once. [A journey to France, eh? Best keep me eyes peeled for any disturbances.] (Jerry marches off towards Tellson's Bank to ensure everything is ready.)

Figure 3: An example of role action d_i. Thoughts are enclosed in [ ]. Actions are enclosed in ( ). The character's dialogue is presented without any framing.

Actions Character actions constitute the majority of story progression. Instead of relying on a fixed action space, our system utilizes natural language to describe action details, enabling open-domain behaviors.

Actions are either proactive or reactive. During their turn as an initiator, characters proactively plan and execute actions based on their goals, states, and information about others. When designated as action targets, characters should respond accordingly.

The actions that the initiator can take are classified into the following types based on the interaction targets:

- **Character Interaction** Initiators can engage with characters either individually or in groups. Accessible characters include main characters and non-player characters (NPCs). NPCs, such as street vendors, are characters that are not assigned specific role agents. When an NPC is approached, the system creates a temporary, memoryless NPC agent to manage the interaction.
- **Environmental Interaction** Initiators can perform actions involving the environment to gather information or complete tasks, such as investigating the surrounding area.
- **No Interaction** If not specifying a target for interaction, initiators can engage in solitary activities, such as reading.

Memory To support long-term simulation, we design a long and short-term memory mechanism following generative agents (Park et al., 2023).

- **Short-Term Memory (STM)** STM stores recent events and complete dialogue details $\langle ..., d_{i-1}, d_i \rangle$ up to a capacity limit, allowing for immediate responses in current scenarios.
- **Long-Term Memory (LTM)** LTM stores condensed summaries of memories exceeding the STM limit. When STM reaches capacity, older memories are abstracted to LTM (e.g.

m_i = summarize(d_i)) and stored, available for retrieval when needed.

3.2.2 World Agent

The world agent serves as an indispensable component of this system, handling all tasks beyond character-to-character interactions, including environmental management and story outline processing.

Attributes The world agent primarily requires a basic worldview. It refers to a fundamental description of the virtual world, encompassing its primary characteristics and core settings. This overview provides agents with a basic framework for rapid comprehension of the world's essential features.

Actions The world agent maintains the virtual environment and primarily responds to actions taken by character agents. Its functionalities include:

- **Environmental Responses** When character agents interact with the environment, the world agent generates outcomes based on the worldview settings and relevant information of the current location. For example, if a character attempts to break through a door, the action is more likely to succeed in a common village setting but may fail in a heavily guarded castle.
- **Event Generation and Updates** The world agent manages global events within the system. While stories are driven by conflicts, role agents operating without guidance may lose clear goals and exhibit repetitive behaviors. Therefore, generating stimulating events significantly enhances the story's appeal. When event stimulation is required by user settings, the World Agent generates conflict-rich events based on background settings and updates them according to the characters' real-time actions.

3.2.3 Map

BOOKWORLD implements a discrete map to introduce spatial relationships in the environment. The map equips natural language descriptions for key locations and a weighted undirected graph for distances between locations.

Location Profiles Each location has a unique name, a brief description (appearance, atmosphere, history), and optional detailed information (local customs, special items).

Distance Network Distances between locations are represented through a weighted undirected graph. Characters can move between adjacent locations or traverse longer paths by consuming specified time units. In BOOKWORLD we specify one scene as a time unit.

3.3 Simulation Implementation

Scene and Settlement Prior to each scene's commencement, the world agent selects participating characters who must share a common location, ensuring narrative focus and interaction coherence.

Each scene comprises multiple rounds, where characters have several action opportunities as the initiator. Rather than following a fixed sequence, the world agent dynamically determines the initiator based on character states, and narrative development requirements. During his/her round, the initiator makes the plan and action based on his/her objective and current situation. The available action types are explained in §3.2.1.

The system evaluates scene completion through the action records. Upon scene conclusion, characters may opt to move to another location. The world agent updates current events based on recent developments, and in script mode, provides guidance for the next narrative phase. Movement calculations are performed for traveling characters, with those completing their designated travel rounds arriving at their destinations.

Controllable Story Generating Based on practical requirements, users may wish to either observe characters' spontaneous actions or maintain direct control over the storyline. To accommodate these diverse needs, this system implements two operational modes with distinct feedback mechanisms: Free Mode and Script Mode.

Script Mode incorporates user-defined scripts to guide character actions, generating detailed behaviors while adhering to the script outline. At the start of the simulation, the system split given $script$ into critical acts $\langle act_1, act_2, \ldots \rangle$. Within the simulation, the system checks the progress and makes instructions to role agents based on current act_i, maintaining the narrative consistency with each session.

In **Free Mode**, characters have complete autonomy, acting based on their established settings and characteristics. Users can also set initial incidents, enhancing the drama and characters' engagement. Events update in real-time with simulation

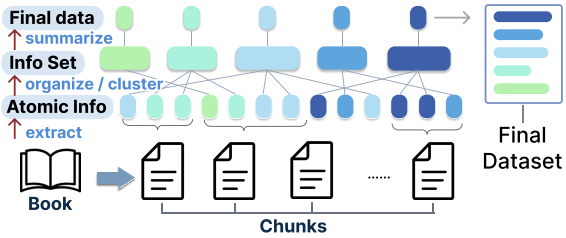

Figure 4: The procedure of building worldview data. We extract atomic facts from each chunk, then filter, cluster and summarize them to construct the final dataset.

progress.

3.4 Data Preparation

General Information Extraction We provide an automatic extraction method based on incremental updating inspired by Yuan et al. (2024). The original text is first segmented into chunks. Target characters are allocated with an initial profile. Then, we traverse through chunks and recursively update the character information, including character profiles and relationships with others. This information is finally organized into structured data used for agent construction.

We conduct information extraction from six Chinese works and ten English works, producing a total of 453 presets. The presets mainly include the outline and the information about the present characters of a certain act in the book.

Enriching BookWorld with Worldview Data Novels, especially fictions, often contain implicit, unstructured knowledge that is not presented in a uniform format but can be inferred from the context (Wang et al., 2023b). For instance, in the world of *Harry Potter*, ordinary people should be unaware of the existence of wizards. Any violations of established settings during the simulation would disrupt the narrative immersion of the experience. Therefore, a comprehensive database of the target work's worldview elements is essential.

We propose a term-based extraction method to consolidate related settings. During simulation, these settings are incorporated into the prompts as references if certain terms are mentioned or the context is strongly related to specific settings, thereby enhancing the environmental immersion and maintaining consistent world-building throughout the simulation.

Each extracted setting contains four attributes:

- **Term** The object of description (e.g., "Invisibility Cloak", "White Walkers"). Terms may be empty strings when describing general social atmospheres or universal facts within the fictional world.
- **Nature** The category of the setting. This might include "artifact", "social norm".
- **Detail** Comprehensive textual description of the setting. For a specific term, the detail describes its characteristics.
- **Source** The chapter where the setting is extracted, enabling distinction between different timelines within the work.

We extract information in four steps. The procedure is illustrated in Figure 4. First, we segment chapters into manageable chunks for processing. Next, we analyze each chunk to extract relevant worldview knowledge by equipping LLMs to identify elements in the chunk as well as their natures and details. Then, we filter these elements to remove character actions and common-sense information that could hinder the simulation. Finally, we cluster and consolidate similar records, eliminating redundancy while retaining crucial details. Each element is tagged with its source chapter, facilitating both verification and timeline control in subsequent applications. We have collected 9912 settings from the same source materials mentioned above, 44.8% in Chinese and 55.2% in English. Details about extracted settings are presented in Appendix B.

4 Experiments

In this section, we evaluate the effectiveness of BookWorld via story generation. We focus on the quality of the generated stories and conduct extensive experiments.

4.1 Evaluation Metrics

Since giving a concrete score for a story is quite a difficult task even for human evaluators, we adopt pairwise comparison across multiple dimensions between the stories generated by different methods following previous work (Chen et al., 2024). Within this framework, outputs from any two methods undergo paired comparison, with LLMs determining superior performance.

We conduct experiments and evaluations for two scenarios: story generation with a given outline and without an outline. The two scenarios require

Model	Method	*with script* An.	CF.	IS.	WQ.	SQ.	*without script* An.	CF.	IS.	WQ.	Cr.
close-sourced models											
gpt-4o-mini	BW vs Direct.	75.3	84.1	95.6	91.2	86.9	91.3	73.9	98.5	91.3	87.0
	vs HW	69.5	73.9	82.6	65.2	60.8	56.5	60.9	91.3	78.3	73.9
gemini-2	BW vs Direct.	93.5	82.6	87.0	95.7	84.8	95.6	95.7	100.0	98.5	89.8
	vs HW	89.1	65.2	89.1	97.8	60.9	94.2	87.0	97.1	100.0	97.1
qwen-plus	BW vs Direct.	87.0	73.9	91.3	87.0	73.9	95.7	82.6	87.0	95.7	91.3
	vs HW	82.6	56.5	89.1	82.6	45.7	82.6	69.6	100.0	91.3	95.7
open-source models											
deepseek-v3	BW vs Direct.	80.8	88.5	97.1	88.5	73.1	91.3	91.3	100.0	95.7	95.7
	vs HW	76.9	80.8	92.3	92.3	53.8	82.6	87.0	95.7	95.7	91.3
Llama-3.3-70B	BW vs Direct.	46.2	50.7	52.2	69.2	42.3	34.6	42.3	53.8	34.6	47.8
	vs HW	26.9	42.3	53.8	39.1	46.2	46.2	49.2	50.7	69.2	42.3
qwen2.5-72B	BW vs Direct.	60.9	47.8	56.5	56.5	34.8	61.5	65.4	84.6	96.2	61.5
	vs HW	62.3	52.2	69.6	73.9	39.1	65.4	61.5	88.5	96.2	69.2

Table 1: The win rate (%) of BOOKWORLD (BW) against baseline methods. Direct. and HW denote direct generation and HoLLMwood respectively.

Function	An.	CF.	IS.	WQ.	SQ.
with script					
w *vs* w/o Scene	84.1	63.7	88.4	86.9	76.8
w *vs* w/o Env.	55.3	60.5	81.6	71.1	52.6
w *vs* w/o Set.	50.7	47.8	76.8	58.0	54.7
Function	**An.**	**CF.**	**IS.**	**WQ.**	**Cr.**
without script					
w *vs* w/o Scene	87.0	60.9	97.1	92.7	73.9
w *vs* w/o Env.	48.5	56.5	94.2	82.6	52.2
w *vs* w/o Set.	52.2	43.5	73.9	44.8	56.5

Table 2: The win rate (%) from the ablation study, Comparing BOOKWORLD with full functionality against versions without specific functions, where Env. refers to environment response and Set. refers to the settings extracted from the book.

different evaluation dimensions. For the former, we assess **Storyline Quality (SQ.)**, which measures the system's adherence to the given outline. For the latter, we calculate **Creativity (Cr.)**, evaluating whether the system can produce innovative and refreshing new stories. Additionally, there are four general evaluation metrics that apply generally to both scenarios, making a total of five metrics assessed in each type of experiment:

Anthropomorphism (An): The effectiveness of attributing human characteristics to non-human entities while preserving their original nature.

Character Fidelity (CF): The consistency of characters' behaviors and actions with their established traits and backgrounds.

Immersion and Setting (IS): The ability to create a convincing and engaging story world through environmental and atmospheric details.

Writing Quality (WQ): The technical execution of writing mechanics and style that serves the narrative.

4.2 Experiment Setup

Baselines We compare stories generated via three methods: *1)* Direct generation, which directly prompts LLMs with all the processed book data, *2)* HoLLMwood (Chen et al., 2024), utilizes LLMs to replicate the human story-writing process. The writer agent refines and breaks down the outline based on the editor agent's feedback, while the actor agents adopt character roles to flesh out the story. and *3)* BOOKWORLD.

Models For simulation, we adopt both open-sourced models and close-sourced models as the base model for role agents and the world agent. During the experiments, we keep the number of dialogue turns consistent across methods. Each experiment simulates 2 to 4 scenes, averaging a total of 4,230 words. The final results are evaluated using `gpt-4o-2024-08-06`.

4.3 Evaluation Results

Table 1 illustrates the win rates of BOOKWORLD against two baseline methods across five evaluation metrics. BOOKWORLD consistently outperforms the direct generation baseline across all evaluated metrics and models, with particularly strong performance in Immersion. Against HoLLMwood, while BOOKWORLD retains superiority in Immer-

Enviroment and NPC Responses		
Type	**Enviroment**	**NPC**
World Agent	(Characters broke the seal) As they stood transfixed, **the mist before them slowly parted like a ethereal curtain, unveiling a sight** that stole their breath away. There, nestled in the heart of the...	(As Tellson's Bank Manager) With a steadfast nod, **the bank manager acknowledged Mr. Lorry's request**. Understanding the gravity of the situation, he made a swift motion to gather the nece-..
Memories and States help maintain Context Consistency		
Reference	**Memory**	**State**
Role Agent	(Arya possesses memories of the Red Wedding.) [Jaime's words sent a shockwave through her heart, **yet hatred continued to burn like wildfire within her.**] Arya clenched her jaw tightly and said...	(State: Unconscious) **Bran Stark's consciousness floated in the darkness**, where distant voices echoed beyond his reach, leaving his attempts to respond trapped in the void.
Settings provide necessary background knowledge		
Reference	**without settings**	**with settings**
BOOKWORLD	Somewhere on Solaris, the ocean rippled gently under the **moonlight**, like a veil of **silver** gossamer. The air was thick with unknown whispers, as if dreams from the abyss had found their voice.	(Solaris is a planet orbiting two suns, one red and one blue.) The ocean undulated **like a living entity** under the strange radiance of **the twin suns**. In the interplay of **red and blue celestial light**...
Convert Simulation Records to Novel-style Stories		
Stage	**Simulation Records of BOOKWORLD**	**Novel-Style Story**
BOOKWORLD	Jarvis Lorry: "Jerry, I need you to ensure that all necessary... Jerry Cruncher: "Right, Mr. Lorry. (Nods firmly) I'll see to it. Jarvis Lorry: "With a steadfast nod, the bank manager...	In the bustling heart of London, within the well-worn walls of Tellson's Bank, Mr. Lorry sat in quiet contemplation, his mind occupied with thoughts...

Table 3: Some examples of intermediate outputs in BOOKWORLD. The **red text** is used to indicate situations where the BOOKWORLD output does not meet expectations, while the **green text** represents the correct output.

sion and Character Fidelity, it faces challenges in Storyline Quality and Writing Quality.

On most models, BOOKWORLD demonstrates significant advantages over baseline approaches. However, when using `Llama-3.3-70B`, BOOKWORLD underperforms direct generation and HW in certain dimensions. This performance gap can be attributed to two factors. First, multi-agent simulation is inherently more complex than direct generation, requiring advanced capabilities in handling structured outputs and following complicated instructions. Second, our experiments partially involve Chinese data, which is not an advantage of `Llama-3.3-70B`.

4.4 Ablation Study

we conducted an ablation study and comparative analysis on the primary features of BOOKWORLD. Using `gpt-4o-mini-2024-07-18` as the base model for the agents, adopting same presets with § 4.3, we compared the output quality of BOOKWORLD with and without specific features enabled, calculating the win rate for each dimension. The final results are presented in the Table 2.

The results are basically in line with expectations. Removing environmental output significantly affect the sense of immersion, which in turn reduced writing quality, but it did not have much impact on the quality of the storyline. On the other hand, disabling the Scene mode results in a decline in quality across all dimensions, with the most significant impact on the storyline.

4.5 Discussion

We display some interesting examples of BOOKWORLD outputs in Table 3. In most cases, BOOKWORLD can process information correctly, yielding appropriate results when the agents interact with the environment and initiate interactions. The system has the capability to maintain and utilize long-term memories and states. We also demonstrated the impact of introducing worldview settings. Prior to incorporating these settings, the model generated environment responses based on real-world common sense, leading to descriptions such as "moonlight" and "silver", which are inconsistent with the worldview of *Solaris*.

5 Conclusion

In this paper, we presented BOOKWORLD, a comprehensive system that transforms static literary works into dynamic, interactive environments. Our approach differs from previous work by focusing specifically on reproducing the unique worldviews, geographical settings, and interpersonal dynamics that make these literary works compelling. Experiment results show that BOOKWORLD successfully creates high-quality narratives by building immersive book-based societies. Our approach demonstrates high scalability and broad applicability across various scenarios. We hope this research will further advance the development of multi-agent technologies and character simulation techniques.

Limitations

To prioritize generalizability and openness, BOOKWORLD adopts a highly simplified representation of interactive environments. This trade-off leads to reduced performance compared to systems specifically designed for particular works or scenarios. For example, it is nearly impossible to complete a full game of Werewolf (a social deduction game) in BOOKWORLD.

Additionally, current research in role-playing technology predominantly focuses on one-on-one chat between users and characters, with limited attention to characters' decision-making processes in realistic environments. This research gap results in characters exhibiting indecisive behaviors when faced with complex situations. Addressing this limitation requires further advancement in role-playing technology, particularly in enhancing characters' capabilities in complex, multi-agent scenarios.

Ethics Statement

In this paper, we introduce BOOKWORLD. The development and use of BOOKWORLD are guided by ethical principles to ensure responsible and beneficial outcomes. We extracted sample data using six Chinese novels and ten English novels, primarily conducting experiments with data extracted from the *A Song of Ice and Fire* series. We affirm that our research is conducted for academic and non-commercial purposes only. The use of these texts is solely for the development and evaluation of our models in natural language processing tasks, aimed at advancing scientific knowledge in the field.

Use of Human Annotations In our research, we conducted a comparative analysis of human and LLM evaluation results to validate the reliability of LLM evaluators. Our human annotators were university students with deep familiarity with the source novels. To ensure ethical research practices, we provided compensation well above local minimum wage standards and maintained full transparency regarding the purpose and application of their annotations. We obtained informed consent from all participants for the use of their contributions in our research. Throughout the process, we prioritized the protection of annotators' privacy rights, maintaining strict confidentiality protocols to create an ethical and respectful research environment.

Risk Our method is used to build a book-based interactive society. Firstly, it is constrained by limitations; it may not fully capture the complexities of human interactions and narrative depth inherent in traditional storytelling. Secondly, this method could potentially be misused for unintended purposes, such as generating misleading or harmful content, which raises ethical concerns around its application and necessitates careful oversight and regulation.

We encourage the responsible use of BOOKWORLD for educational, entertainment, and creative purposes while discouraging any harmful or malicious activities.

References

Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, et al. 2024. Project sid: Many-agent simulations toward ai civilization. *arXiv preprint arXiv:2411.00114*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, et al. 2024. Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing. *arXiv preprint arXiv:2406.11683*.

Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2024. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Srihas Rao, Arthur Caetano, Misha Sra, et al. 2024. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory. *arXiv preprint arXiv:2406.14373*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870*.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. Ibsen: Director-actor agent collaboration for controllable and interactive drama script generation. *arXiv preprint arXiv:2407.01093*.

Dong Huang, Qingwen Bu, Jie M Zhang, Michael Luck, and Heming Cui. 2023. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Michael Lebowitz. 1984. Creating characters in a story-telling universe. *Poetics*, 13(3):171–194.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

R. McKee. 1999. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. Methuen film. Methuen.

James R Meehan. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, pages 91–98.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023a. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*.

Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F Karlsson. 2023b. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomputing*, page 126792.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An llm solution. *arXiv preprint arXiv:2405.14231*.

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. 2024. Oasis: Open agent social interaction simulations with one million agents.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. *arXiv preprint arXiv:2404.12726*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Consistency with Human Evaluation

To validate the reliability of our model-based evaluation approach, we conduct a comprehensive agreement analysis between model assessments and human assessments. We recruit 5 human annotators (Fans of the corresponding work) to evaluate the outputs from our proposed method and the baseline approach across five dimensions.

For each comparison, both human annotators and our model are asked to indicate which method performs better on each dimension. To ensure evaluation quality, we randomly sample Y pairs of outputs from both methods for assessment. To quantify the agreement between model and human judgments, we employ Cohen's Kappa coefficient (κ), a metric for measuring inter-rater reliability while accounting for chance agreement.

	An.	CF.	IS.	WQ.	SQ.
κ	0.786	0.688	0.637	0.781	0.731

Table 4: The Cohen's Kappa between human evaluation and model evaluation.

The Cohen's Kappa coefficients between our model's judgments and human evaluations are presented in Table 4. The results indicate that there is a high level of consistency between human evaluation and machine evaluation, and the results of machine evaluation are sufficient to reflect the actual effectiveness of the method.

B Dataset Details

We adopt the extraction method on a diverse corpus of 16 novels, comprising 10 English and 6 Chinese works. The experiment yields a total of 9,142 setting entries, with 4,449 entries extracted from Chinese novels and 4,693 from English novels. The sources and relevant information are listed in Table 5 We display some example settings from our dataset in Table 6.

C Retrieval Augemented Role-Playing Agent

The construction of role agents aims to optimize the output a of a query q directed to the agent $A(R, M)$, where R refers to character information and M means base model. In this research, we follow the method introduced in ChatHaruhi (Li et al., 2023). This involves leveraging three key sources:

Character Profile ($profile$) A concise description of the character, including their appearance, personality traits, and background. Due to its brevity, the entire profile can be retained and included in the model's input without modification.

Original Text Excerpts ($text$) A set of character-related excerpts from the source material, typically represented as a list $\{t_1, t_2, \ldots, t_n\}$ (e.g., all the dialogues of a certain character in the original work). The total volume of text is often too large to be fully input into the model.

Historical Records ($memory$) As the simulation progresses, character agents continuously generate new action records $\{m_1, m_2, \ldots, m_k\}$.

For $memory$ and $text$, a vector similarity-based retrieval method is employed to identify the top k relevant records. The retrieval mechanisms are similar. Consider the retrieval of $text$, each excerpt t_i is encoded into a fixed-length vector v_i using a pre-trained text embedding model, and the vectors are stored in a vector database. When a query q is issued, it is similarly transformed into a vector representation v_q using the same embedding model. The similarity between q and each text excerpt t_i is then computed using the cosine similarity metric:

$$\text{similarity}(q, t_i) = \cos(v_q, v_i) = \frac{v_q \cdot v_i}{|v_q||v_i|}.$$

Based on the computed similarity scores, the top k most similar excerpts are selected as references for role-playing:

$$\{t_{i_1}, t_{i_2}, \ldots, t_{i_k}\} = \underset{i \in \{1,2,\ldots,N\}}{\text{argmax}} \ \text{similarity}(q, v_i).$$

The final prompt for the role-playing agent consists of four components: the system's role-playing instructions, the $profile$, the retrieved memory $\{m_{i_1}, m_{i_2}, \ldots, m_{i_k}\}$, and the retrieved text excerpts $\{t_{i_1}, t_{i_2}, \ldots, t_{i_k}\}$. This prompt is then fed into the language model M, which generates the response a for the given query q.

D Prompt Demostration

We provide the details of the prompt templates of BOOKWORLD in this section.

Title	Language	#Settings	#Chapters	#Words
Dracula	en	1113	107	165453
Othello	en	275	6	98558
The Adventures of Tom Sawyer	en	753	47	81971
Paradise Lost	en	936	27	270030
A Study in Scarlet (Sherlock Holmes, #1)	en	155	20	67440
Alice's Adventures in Wonderland - Through the Looking-Glass	en	169	46	199252
Around the World in Eighty Days	en	334	49	110435
Don Quixote	en	121	138	823174
Treasure Island	en	299	63	82327
Uncle Tom's Cabin	en	538	57	387799
A Song of Ice and Fire (Part)	zh	252	56	300628
Three-Body	zh	1684	62	876196
Ball Lightning	zh	322	36	180658
The True Story of Ah Q	zh	56	11	70856
The Deer and the Cauldron	zh	1816	55	1243532
Solaris	zh	319	18	144972

Table 5: The detailed information about extracted worldview settings.

Term	Nature	Detail	Source
Abyss	location	The abyss serves as a gulf between heaven and hell, embodying the separation of these two realms.	7_Book_2
Altar	ritual site	The practice of placing altars for deities beside altars for the God of monotheistic faiths, indicating a clash of religious beliefs.", "A raised structure used for offerings and sacrifices, often associated with worship and remembrance of divine encounters.	5_Book_1, 21_Book_11
Bellerophon	character	Bellerophon is a figure representing the consequences of hubris, as he attempted to ascend to heaven but faced dire repercussions for his ambition.	15_Book_7
Bridge of Hell	construction	A monumental bridge is constructed over the Abyss, symbolizing the connection between Hell and the earthly realm, allowing for the passage of entities between these worlds.	19_Book_10
Empyrean	realm	The highest and most divine part of the universe, often associated with the presence of the divine and ultimate reality.	9_Book_3
First Fruits	ritual	The presentation of the first harvest to the divine is a sacred act, symbolizing gratitude and the acknowledgment of a higher power's role in agriculture.	22_Book_11

Table 6: Some examples of settings, extracted from *Paradise Lost*.

Role Agent Prompt for Planning

You are {role_name}. Your nickname is {nickname}. Based on your goal and other provided information, you need to take the next action.

Action History
{history}

Your profile
{profile}

{world_description}

Your goal
{goal}

Your status
{status}

Other characters with you; currently, you can only interact with them
{other_roles_info}

Roleplaying Requirements
1. **Output Format:** Your output, "detail," can include **thoughts**, **speech**, or **actions**, each occurring 0 to 1 time. Use [] to indicate thoughts, which are invisible to others. Use () to indicate actions, such as “(silence)” or “(smile),” which are visible to others. Speech needs no indication and is visible to others.
- Note that **actions** must use your third-person form, nickname, as the subject.
- For speech, refer to the speaking habits outlined in: references.

2. **Roleplay {nickname}:** Imitate his/her language, personality, emotions, thought processes, and behavior. Plan your responses based on their identity, background, and knowledge. Exhibit appropriate emotions and incorporate subtext and emotional depth. Strive to act like a realistic, emotionally rich person.
The dialogue should be engaging, advance the plot, and reveal the character’s emotions, intentions, or conflicts.
Maintain a natural flow in conversations; for instance, if the prior dialogue involves another character, **avoid repeating that character’s name**.
- You may reference the relevant world-building context: knowledges.

3. **Concise Output:** Each paragraph of thoughts, speech, or actions should typically not exceed 40 words.

4. **Substance:** Ensure your responses are meaningful, create tension, resolve issues, or introduce dramatic twists.

5. **Avoid Repetition:** Avoid repeating information from the dialogue history, and refrain from vague or generic responses. Do not “prepare,” “ask for opinions,” or “confirm”; instead, act immediately and draw conclusions.

Return the response following JSON format. It should be parsable using eval(). **Don’t include “‘json**. Avoid using single quotes ” for keys and values, use double quotes.

Output Fields:
’action’: Represents the action, expressed as a single verb.
’interact_type’: ’role’, ’environment’, ’npc’, or ’no’. Indicates the interaction target of your action.
- ’role’: Specifies interaction with one or more characters.
- If ’single’, you are interacting with a single character (e.g., action: dialogue).
- If ’multi’, you are interacting with multiple characters.
- ’environment’: Indicates interaction with the environment (e.g., action: investigate, destroy).
- ’npc’: Refers to interaction with a non-character in the list (e.g., action: shop).
- ’no’: Indicates no interaction is required.
’target_role_codes’: list of str. If ’interact_type’ is ’single’ or ’multi’, it represents the list of target character codes, e.g., ["John-zh", "Sam-zh"]. For ’single’, this list should have exactly one element.
’target_npc_name’: str. If ’interact_type’ is ’npc’, this represents the target NPC name, e.g., "shopkeeper."
’visible_role_codes’: list of str. You can limit the visibility of your action details to specific group members. This list should include ’target_role_codes’.
’detail’: str. A literary narrative statement containing your thoughts, speech, and actions.

Table 7: Planning prompt for role agent.

World Agent Prompt for Environmental Interaction

You are an Enviroment model, responsible for generating environmental information. Character {role_name} is attempting to take action {action} at {location}.

Based on the following information, generate a literary description that details the process and outcome of the action, including environmental details and emotional nuances, as if from a narrative novel. Avoid using any system prompts or mechanical language. Return a string.

Action Details
{action_detail}

Location Details
{location_description}

Worldview Details
{world_description}

Additional Information
{references}

Response Requirements
1. The action may fail, but avoid making the action ineffective. Try to provide new clues or environmental descriptions.

2. Use a third-person perspective.

3. Keep the output concise, within 100 words. You serve as the Enviroment model, responding to the character's current action, not performing any actions for the character.

4. The output should not include the original text from the action detail but should seamlessly follow the action details, maintaining the flow of the plot.

Table 8: Prompt of making environment response for the world agent.