EvolvingWorld: An Open-Schema Framework for Co-Evolving Role-Play Agents and World Model in Interactive Literary World

Qing Zong[1], Yue Guo[2], Mengxin Yang[3], Yiwen Guo[4], Yangqiu Song[1]
[1]Hong Kong University of Science and Technology [2]LIGHTSPEED
[3]Huazhong University of Science and Technology [4]Independent Researcher
qzong@cse.ust.hk

Abstract

This paper introduces **EvolvingWorld**, a framework and benchmark for character and world co-evolution in interactive literary worlds. Existing systems either treat interactive literary simulation as static persona imitation or isolated scene generation, failing to capture how characters and worlds evolve together over time. To address this, EvolvingWorld models literary simulation as a long-horizon process where characters interact, scenes progress, and character and world states are persistently updated. Unlike prior systems relying on fixed schemas, EvolvingWorld adopts an open-schema framework to support simulation across diverse literary worlds. The framework consists of two coupled modules: a **Character Agent** for multi-character role-play and persistent profile evolution, and an LLM-based **World Model** for global and location/entity-level state maintenance and scene progression. Based on this architecture, we formulate 7 trainable tasks for scene initialization, interaction generation, and state update. We construct a dataset from 57 books, producing 138,596 supervised training samples and 222 snapshots for testing. Furthermore, we introduce a trajectory-level LLM-as-Judge evaluation protocol spanning 10 dimensions and 20 metrics. Experiments show that EvolvingWorld can improve long-horizon simulation by effectively maintaining persistent, coherent character and world development. [1]

1 Introduction

Large language models (LLMs) have enabled fluent role-playing agents that imitate fictional characters and sustain persona-grounded dialogue (Shao et al., 2023; Wang et al., 2024, 2025b; Zhou et al., 2024; Xu et al., 2026a,b). Yet simulating a literary world poses a harder long-horizon challenge: as a story unfolds, characters revise beliefs, motivations, and relationships, while locations, objects, and background conditions also change. The goal is therefore not only to produce the next plausible utterance, but to maintain coherent character and world states across scenes.

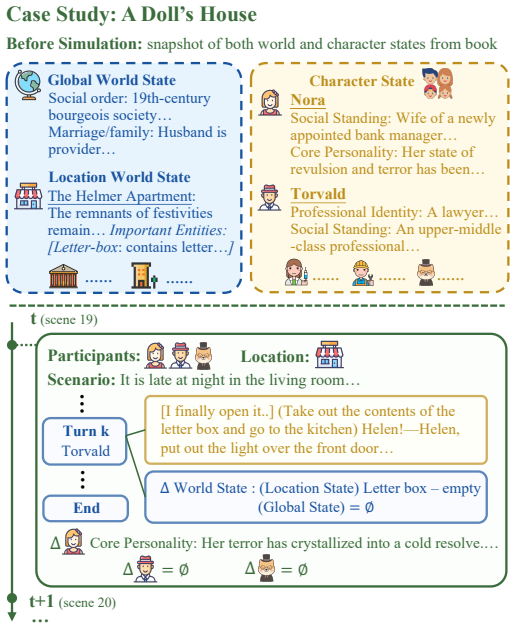


Figure 1: A simulation case from EvolvingWorld. Starting from a snapshot derived from a book, multiple characters interact as both character and world states evolve.

Existing role-playing systems fall short in three ways. First, most persona-based agents rely only on static profiles or short dialogue contexts (Li et al., 2023; Lu et al., 2024; Yang et al., 2025; Zhou et al., 2025; Liu et al., 2026b). Second, multi-agent environments often use manually specified sandboxes, making them hard to scale to diverse literary worlds (Park et al., 2023; Yan et al., 2023; Yu et al., 2025b). Third, prior book-grounded systems focus on single-scene role-play (Wang et al., 2025b; Xu et al., 2026b) or partial long-horizon updates. For example, BookWorld (Ran et al., 2025) updates characters' goals and states, and global events, but lacks full profile evolution, location/entity-level world updates, and trainable subtask supervision.

We argue that literary world simulation requires

[1]https://github.com/HKUST-KnowComp/EvolvingWorld

Framework	Character			Interaction		Message Components					World					Data
	Profile	Prof. Upd.	Open-Sch.	Multi-Char.	Init Scene	Speech	Thought	Action	Env.	Group Act	Multi-W.	Global	Glob. Upd.	Open-Sch.	Loc./Ent.	Training
ChatHaruhi (Li et al., 2023)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
CharacterLLM (Shao et al., 2023)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
HPD (Chen et al., 2023)	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
RoleLLM (Wang et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
CharacterGLM (Zhou et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓
CharacterEval (Tu et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓
CharacterBench (Zhou et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
MMRole (Dai et al., 2025)	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
DITTO (Lu et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
SimsChat (Yang et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
BeyondDialogue (Yu et al., 2025a)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
Crab (He et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓
TailorRPA (Gao et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
CoSER (Wang et al., 2025b)	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
AdaMARP (Xu et al., 2026b)	✓	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
BookWorld (Ran et al., 2025)	✓	●	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	●	✗	●	✗
GenerativeAgents (Park et al., 2023)	✓	●	✗	✓	●	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗
LARP (Yan et al., 2023)	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓
CharacterBox (Wang et al., 2025a)	✓	●	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	●	✗
EvolvingWorld (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison with representative role-play frameworks. Character columns cover explicit profiles, automatic profile updates, and whether character states use open rather than fixed schema. Interaction columns cover multi-character scenes and automatic scene initialization or switching. Message columns indicate whether generated messages include speech, thought, action, environmental messages, and collective messages by multiple characters. World columns cover support for diverse worlds, global world settings, global world state evolution, open-schema world representations, and location/entity-level state modeling. Symbols indicate full support (✓), partial support (●), or no support (✗). Detailed clarifications for partial support are provided in Appendix A.

open-schema co-evolution. Open-schema means that the system infers the relevant character and world dimensions from each book rather than forcing all stories into fixed slots: character dimensions can differ between a detective's investigative habits and a Victorian orphan's social position, while world dimensions may shift from school rules and class hierarchies to political orders or supernatural systems. Once these dimensions are constructed, co-evolution keeps character and world states coupled: character actions can reshape locations or social orders, while world changes can in turn alter motivations and profiles. The key difficulty is deciding which dimensions to track, when evidence justifies profile updates, and how local events propagate to global, location, and entity states. Thus, we reframe interactive literary world from characters and worlds that remain largely static to ones that can fully develop beyond their initial descriptions as long as supported by ongoing interactions.

We present **EvolvingWorld**, a framework for **open-schema** character and world **evolution** in interactive multi-agent literary worlds. Given a book snapshot, EvolvingWorld simulates the story forward with two coupled modules: an open-schema **Character Agent** for multi-character role-play and profile evolution, and an LLM-based **World Model** for global, location, and entity-level state tracking and scene progression (Figure 1). Within the Character Agent, profile dimensions may evolve at different speeds: emotions can shift quickly, while personality traits often require accumulated evidence. We therefore use a **hidden tracker** to store weak or emerging evidence separately before they justify profile updates. We decompose the framework into 7 supervised tasks covering scene initialization, interaction generation, and state update.

We construct a dataset from 57 books, yielding **138,596** supervised training samples and **222** test snapshots. We also introduce a trajectory-level LLM-as-Judge evaluation framework covering 20 metrics, measuring persistent character and world development beyond local role-playing quality. Experiments show that EvolvingWorld reduces the long-horizon performance degradation often observed in previous frameworks (Figure 3).

Overall, we make 4 contributions: **(1)** a new formulation of interactive literary world simulation as a long-horizon process in which characters and the world can fully develop beyond their initial descriptions based on interactions rather than updating only a few predefined fields, shifting the objective from reproducing the original book toward sustaining a grounded, continuously developing world; **(2)** EvolvingWorld, a simulation framework that couples an open-schema Character Agent with hidden trackers for multi-timescale profile evolution and an LLM-based World Model that maintains open-schema global and location/entity-level states while guiding scene progression, decomposed into 7 trainable tasks; **(3)** a benchmark built from 57 books with **138,596** training samples and **222** test snapshots, together with trajectory-level evaluation framework covering 20 metrics; and **(4)** empirical evidence that EvolvingWorld reduces long-horizon performance degradation, demonstrating the value of co-evolving character and world states.

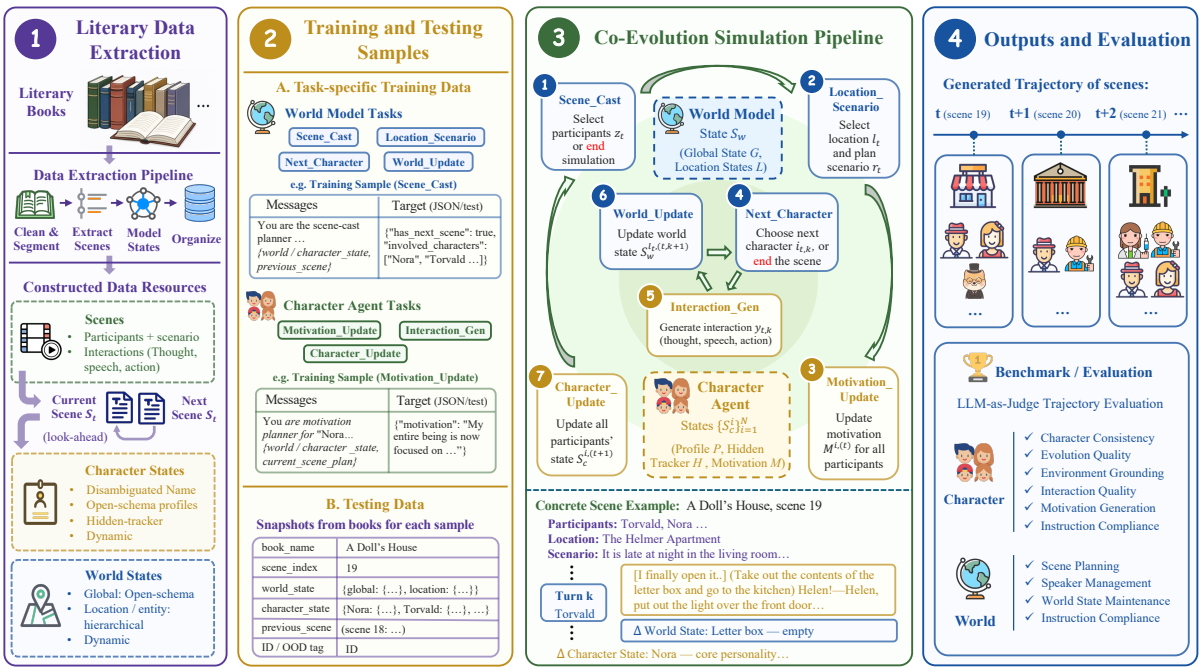


Figure 2: Overview of our EvolvingWorld framework. It contains following stages: Dataset Construction, parts 1-2 (Section 3.3); Simulation Pipeline, part 3 (Section 3.2); and Evaluation Method, part 4 (Section 3.4).

2 Related Work

Role-Play Agents. Role-playing language agents have progressed from scripts (Wang et al., 2024) and self-alignment (Lu et al., 2024; Yang et al., 2025) to training-based personality simulation (Shao et al., 2023; Zhou et al., 2024; Wang et al., 2025c), multi-modality (Dai et al., 2025) and memory retrieval (Gao et al., 2025; Yu et al., 2025a). While these systems typically treat personas as static anchors, we enables open-schema character evolution over long-horizon narratives.

Multi-Agent and World Model. Believable simulation needs grounded interaction. While social (Park et al., 2023; Piao et al., 2026; Zhang et al., 2025) and gaming (Yan et al., 2023; Wang et al., 2025a) environments rely on fixed sandboxes, which do not scale to diverse worlds, role-play agents like CoSER (Wang et al., 2025b) and AdaMARP (Xu et al., 2026b) lack world modeling. BookWorld (Ran et al., 2025) contains a world agent but uses a predefined fixed schema and only static world states. Recent advances explore LLM-based world models (Chu et al., 2026; Li et al., 2026) and also open-schema in event extraction (Bai et al., 2026; Lu et al., 2026), which haven't been studied in interactive literary worlds. We bridge this gap by supporting LLM-based open-schema world evolutions for diverse book worlds.

Role-Play Evaluation. Evaluation has shifted from fluency toward behavior (Tu et al., 2024; Zhou et al., 2025; Boudouri et al., 2025). While some recent works have introduced trajectory-level evaluation (Xu et al., 2026b; Ye et al., 2025), the quality of state evolution remains underexplored. EvolvingWorld introduces a trajectory-level framework to further quantify persistent character and world development in open-ended environments.

3 The EvolvingWorld Framework

Figure 2 provides an overview of EvolvingWorld, with data construction, simulation pipeline, and evaluation method. It unfolds a book from a chosen narrative point into an evolving interactive literary world. Given a structured state snapshot at that point, it simulates scene-by-scene interactions over persistent character and world states modeling.

3.1 Design Principles of EvolvingWorld

Table 1 compares EvolvingWorld with existing role-play frameworks. Most prior systems face two main limitations: their characters rely on fixed, static profiles that cannot evolve over time; and their worlds are either rigid sandboxes or lack detailed entity-level state tracking. To address these gaps, EvolvingWorld introduces two core components, a **Character Agent** and a **World Model**, to drive open-schema, long-horizon evolution.

3.1.1 Character Agent

The Character Agent represents each character with an open-schema profile, since characters from dif-

ferent books vary a lot: a detective may have investigative habits, while a Victorian orphan is better characterized by social position. Prior literary role-play frameworks (Wang et al., 2025b; Xu et al., 2026b; Ran et al., 2025) rely on fixed profile dimensions. In contrast, we provide only reference dimensions when constructing profiles, allowing LLM to select, merge, or introduce new fields according to the book's genre, setting, and style.

Beyond individual characters, the Character Agent also supports the environment and character groups as special acting units. They enable environmental events and shared group interactions within the same interaction loop.

The Character Agent further supports persistent profile evolution. Unlike prior works (Ran et al., 2025; Park et al., 2023) updating only several predefined dimensions such as memory or psychological state, we treat every dimension in the open-schema profile as a potentially evolvable part of the character state. Moreover, profile dimensions may evolve at different speeds: mood may change quickly, but personality requires accumulated evidence. To model this multi-timescale evolution, we introduce a hidden tracker that records weak or emerging evidence separately from the profile. This design prevents premature profile updates while allowing repeated signals across scenes to accumulate into later changes. During evolution, the Character Agent considers both dimension-level changeability and accumulated hidden evidence before updating states.

3.1.2 World Model

The World Model maintains both a global world state and location-level physical states. The global state captures world-level settings, such as historical background and social institutions. Since literary works construct vastly different worlds, ranging from school narratives to post-apocalyptic societies, we adopt an open-schema design that avoids reducing the global state to fixed dimensions, unlike prior systems like BookWorld (Ran et al., 2025).

In addition to global state, we explicitly model the physical state of each location. Sandbox-based environments such as Generative Agents (Park et al., 2023) and LARP (Yan et al., 2023) depend on a manually predefined single world, which limits scalability across diverse worlds. We instead use an LLM-based World Model to construct detailed physical states for all locations and update them throughout simulation. Locations can be nested with sub-locations, such as a house with rooms, or separate atomic locations, such as the road outside the house. For each level of location, the framework maintains a detailed description and tracks all important non-character entities and their states, such as a Christmas tree standing by the window. Both global and location-level states are automatically updated through character interactions.

Together, these evolving character and world states enable long-horizon, cross-scene simulation in interactive multi-agent literary worlds, where characters are not constrained to static personas and worlds are not treated as passive backdrops.

3.2 Simulation Pipeline and Task Formulation

We now formulate how Character Agent and World Model are composed into a co-evolution simulation pipeline. Starting from a snapshot of a book, the simulator repeatedly plans a scene, generates multi-character interactions, updates the world during the scene, and revises character states after the scene. This turns the design principles above into a sequence of tasks, supporting end-to-end long-horizon simulation. To keep it clear, we define the states and modules' observations, and present the seven tasks in their execution order below.

States and observations. At scene step t, let $\mathcal{I}$ denote the full character set and $\mathcal{L}$ the set of locations. The simulator maintains a global world state, one location state for each location $\ell \in \mathcal{L}$, and one character state for each character $i \in \mathcal{I}$:

$$S_w^{\ell,(t)} = \big(G^{(t)}, L^{\ell,(t)}\big), \quad \ell \in \mathcal{L},$$
$$S_c^{i,(t)} = \big(P^{i,(t)}, H^{i,(t)}, M^{i,(t)}\big), \quad i \in \mathcal{I}.$$

Here $S_w^{\ell,(t)}$ denotes the world state at location ℓ, while $S_c^{i,(t)}$ denotes the state of character i. In particular, $G^{(t)}$ is the open-schema global world state, and $L^{\ell,(t)}$ stores the description of location ℓ and its important entity states. For character i, $P^{i,(t)}$ is the open-schema profile, $H^{i,(t)}$ is the hidden tracker, and $M^{i,(t)}$ is the scene-level motivation.

The observations are module-specific views constructed from these states. For any character subset $A \subseteq \mathcal{I}$ and location subset $B \subseteq \mathcal{L}$, the World Model observation $O_w^{A,B,(t)}$ contains the global state, the location states in B, and the character states in A. For any character i and location ℓ, the Character Agent observation $O_c^{i,\ell,(t)}$ contains the corresponding location-specific world state and

character state:

$$O_w^{A,B,(t)} = \big(G^{(t)}, \{L^{\ell,(t)}\}_{\ell \in B}, \{S_c^{i,(t)}\}_{i \in A}\big),$$
$$O_c^{i,\ell,(t)} = \big(S_w^{\ell,(t)}, S_c^{i,(t)}\big).$$

Task sequence. Each scene is generated by the following ordered tasks.

Task 1: `scene_cast`. The World Model selects the participating character set $z_t \subseteq \mathcal{I}$ from the full character set:

$$O_w^{\mathcal{I},\varnothing,(t)} \to z_t.$$

Task 2: `location_scenario`. Given the selected cast, the World Model produces a scene plan that specifies the location ℓ_t and scenario r_t:

$$(O_w^{z_t,\mathcal{L},(t)}, z_t) \to \ell_t, r_t.$$

Task 3: `motivation_update`. Before the scene begins, the Character Agent prepares each participant with a scene-specific motivation:

$$(O_c^{i,\ell_t,(t)}, r_t) \to M^{i,(t)}, \quad i \in z_t.$$

Task 4: `next_character`. During the scene, let $Y_{t,<k}$ be the interaction history before turn k. The World Model chooses the next acting character from the selected cast:

$$(O_w^{z_t,\{\ell_t\},(t)}, r_t, Y_{t,<k}) \to i_{t,k}.$$

Task 5: `interaction_gen`. Given the selected actor $i_{t,k}$, which may be a character, the environment, or a character group, the Character Agent generates $y_{t,k}$. Its content may mix thought `[...]`, speech as plain text, and action `(...)`. Only the character's own thoughts in $Y_{t,<k}$ are visible:

$$(O_c^{i_{t,k},\ell_t,(t)}, r_t, Y_{t,<k}) \to y_{t,k}.$$

Task 6: `world_update`. After each interaction, the World Model updates the global and location state:

$$(O_w^{z_t,\{\ell_t\},(t)}, y_{t,k}, Y_{t,<k}) \to S_w^{\ell_t,(t,k+1)}.$$

The completed scene history is denoted by $Y_t = (y_{t,1}, \ldots, y_{t,K_t})$, where K_t is the number of generated turns before the scene ends.

Task 7: `character_update`. After the scene, the final in-scene world state becomes $S_w^{\ell_t,(t+1)}$, and each participating character state is updated:

$$(O_c^{i,\ell_t,(t)}, Y_t, S_w^{\ell_t,(t+1)}) \to S_c^{i,(t+1)}, \quad i \in z_t.$$

Non-participating characters and non-current locations keep their previous states.

Together, these tasks define one scene-level transition over the persistent states and interactions,

$$\big(S_w^{\mathcal{L},(t)}, S_c^{\mathcal{I},(t)}\big) \to \big(S_w^{\mathcal{L},(t+1)}, S_c^{\mathcal{I},(t+1)}, Y_t\big) = \big(O_w^{\mathcal{I},\mathcal{L},(t+1)}, Y_t\big),$$

and repeated transitions produce the full trajectory,

$$\tau = \big(O_w^{\mathcal{I},\mathcal{L},(0)}, z_1, r_1, Y_1, O_w^{\mathcal{I},\mathcal{L},(1)}, \ldots, z_T, r_T, Y_T, O_w^{\mathcal{I},\mathcal{L},(T)}\big).$$

The goal is to produce a grounded long-horizon trajectory in which scene content, character evolution, and world-state changes remain mutually consistent. See Appendix E.1 for pseudocode.

3.3 Dataset Construction

Following Wang et al. (2025b), we select 57 chronologically narrated books. We use *Gemini-2.5-Pro* as extraction LLM. Chronological narratives allow later scenes to serve as look-ahead references, so character and world-state changes are grounded in textual evidence rather than LLM's knowledge. Detailed construction and prompts are provided in Appendices C and K.1.

3.3.1 Structured Data Extraction

EvolvingWorld segments each book into text chunks and uses the extraction LLM to build structured scenes with summaries, scenarios, key characters, and multi-turn interactions. If a scene is cut off at a chunk boundary, the truncated text is prepended to the next chunk, so the LLM can continue the same scene.

For characters, the LLM first unifies different names that refer to the same character from extracted mentions. It then builds each character's initial state from the first few relevant scenes, including an open-schema profile and hidden tracker, and updates this state scene by scene using later narrative evidence as look-ahead. For example, if an idle student reflects after a failure, a later scene showing sustained effort confirms that the reflection led to a real character-state change.

World states follow the same principle. The global world state tracks story-level settings and systemic conditions, while location-level states track each location's description and important non-character entities. The immediately following interactions and scene provide reference evidence for global-state changes, while interactions and the next scene happening at the same location provide evidence for location's physical-state changes. The LLM standardizes location names, initializes both world states, and updates them after interactions.

3.3.2 Train/Test Split Construction

We then split them into training data, in-domain (ID) test data from partially seen books, and out-of-

CC: Character Consistency **EQ**: Evolution Quality **EG**: Environmental Grounding **IQ**: Interaction Quality **MG**: Motivation Generation
IC: Instruction Compliance **PF**: Profile Fidelity **SSF**: Speaking Style Fidelity **MDB**: Motivation-Driven Behavior **PUF**: Profile Update Fidelity
PES: Profile Evolution Smoothness **EA**: Environment Awareness **EU**: Environmental Utilization
CR: Contextual Responsiveness **NP**: Narrative Progression **MQ**: Motivation Quality

Models	CC			EQ		EG		IQ		MG	IC	Avg.
	PF	SSF	MDB	PUF	PES	EA	EU	CR	NP	MQ	IC	
Closed-source												
Kimi-K2.5	80.79±10.84	61.45±19.00	87.52 ±9.78	81.07 ±8.63	54.51±16.38	81.29±10.59	83.86±13.00	88.31 ±8.27	56.02±16.66	77.12±12.25	82.83 ±5.68	75.89
Gemini-2.5-Flash	75.34 ±3.94	61.68 ±6.80	75.67 ±5.96	68.61 ±5.91	58.02 ±7.68	67.02 ±6.14	57.10 ±8.90	75.72 ±4.02	51.86 ±9.17	74.61 ±8.04	51.45±26.93	65.19
Gemini-2.5-Pro	93.88 ±3.47	85.94 ±5.90	94.95 ±3.44	83.28 ±5.33	70.51 ±8.65	89.53 ±5.19	90.12 ±6.15	93.63 ±3.25	72.10 ±8.19	77.20 ±6.36	86.06 ±3.08	85.20
Gemini-3.1-Pro-P	89.79 ±5.23	79.40 ±6.91	93.09 ±3.57	78.90 ±5.87	66.29 ±9.98	85.35 ±6.17	83.90 ±8.33	91.87 ±3.62	67.74 ±9.63	83.04 ±6.41	84.12±12.90	82.14
GPT-4o	71.45 ±6.84	54.13 ±9.15	76.36 ±6.76	57.66 ±8.60	54.46 ±8.62	75.80 ±7.04	76.32 ±9.16	77.94 ±4.63	55.89±11.73	69.94 ±8.81	74.15±16.39	67.65
GPT-5-Chat	80.73 ±5.60	73.44 ±9.14	85.89 ±5.45	71.33 ±7.30	64.00 ±9.68	86.74 ±5.29	92.52 ±6.49	87.36 ±4.09	67.37 ±9.19	81.49 ±7.85	77.29±20.63	78.92
GPT-5.3-Chat	94.71 ±2.75	88.77 ±4.30	93.71 ±2.95	77.90 ±3.74	80.16 ±6.62	91.62 ±4.88	92.10 ±7.54	92.27 ±3.74	59.98 ±7.55	83.94 ±6.50	83.85 ±3.25	85.36
Claude-4.6-Sonnet	96.47 ±4.16	95.52 ±4.51	98.49 ±3.29	95.30 ±5.61	71.02±15.19	94.33 ±5.25	96.79 ±7.27	97.29 ±4.67	84.03 ±9.31	90.17 ±5.07	62.07±35.76	89.23
Claude-4.6-Opus	**97.81** ±1.92	**96.70** ±2.79	**99.00** ±1.63	**96.36** ±2.83	**84.62** ±9.15	**95.91** ±3.36	**98.73** ±2.46	**98.85** ±1.49	**89.76** ±6.06	**95.36** ±3.63	**91.53** ±3.25	**94.97**
Open-source												
< 14B												
Qwen3-4B-I	15.31±10.34	7.25 ±6.89	13.57±10.80	21.84±10.54	26.89±11.14	30.11±11.88	32.68±10.21	15.69±12.49	12.47 ±9.73	39.00±18.49	24.76±17.68	21.77
Qwen2.5-7B-I	15.77 ±6.68	7.69 ±5.06	11.70 ±6.98	32.25 ±7.42	25.07±10.49	30.45 ±6.41	28.53 ±6.75	11.80 ±7.31	9.69 ±5.53	36.20±16.06	20.14±12.72	20.84
Qwen-7B (Coser)	16.59±11.97	12.20 ±8.33	27.61±12.64	18.16±16.58	17.91±16.25	37.34 ±8.69	25.20 ±4.14	13.92 ±6.96	27.60±12.86	2.06 ±9.94	10.26 ±1.55	18.98
Qwen-7B (Crab)	14.25 ±7.83	6.34 ±4.43	15.76 ±7.80	19.82 ±7.22	15.89 ±9.94	32.79±10.41	20.28 ±5.85	16.91 ±8.45	17.74 ±5.31	26.82±12.05	14.91 ±8.10	18.32
Llama-3.1-8B-I	30.58±11.10	22.30 ±8.36	30.71±12.65	21.87±16.04	20.36±13.89	44.33 ±8.88	**46.19** ±9.30	40.78±12.36	**40.77** ±9.04	27.26±22.46	14.55±17.16	30.88
Llama-8B (Coser)	30.36±16.76	22.26±15.53	38.28±18.63	15.86±16.21	15.94±15.91	37.61 ±8.57	31.04 ±2.85	30.36±14.55	37.27±11.12	4.07±14.52	11.09 ±4.60	24.92
Llama-8B (Crab)	29.37±18.52	20.00±16.45	38.10±20.05	1.26 ±5.04	2.27 ±8.62	33.20 ±5.08	32.67 ±2.56	26.57±12.04	34.03 ±7.19	10.45±20.33	10.25 ±1.88	21.65
Qwen-4B (EW-ours)	44.68±10.21	37.24 ±9.73	37.10 ±8.45	35.24 ±7.81	31.43±10.20	40.09 ±6.84	28.15 ±5.41	35.64 ±9.28	28.30 ±5.88	40.57 ±9.72	59.12±11.15	37.96
Qwen-7B (EW-ours)	**54.07** ±9.85	**48.19**±10.44	**47.15** ±8.44	41.24 ±7.72	40.33±10.06	45.44 ±7.27	31.81 ±6.14	**48.61** ±9.92	36.09 ±6.68	42.12 ±8.31	**65.78** ±9.45	45.53
Llama-8B (EW-ours)	50.94±20.32	44.90±18.45	46.43±16.41	**42.49**±10.11	**42.44**±10.66	**46.47**±11.92	33.75 ±7.68	46.03±19.69	38.83±13.13	**47.96** ±7.72	65.62 ±8.79	**45.99**
≥ 14B												
Qwen2.5-14B-I	31.44±10.64	14.71 ±7.63	30.43±10.11	43.95 ±6.51	35.56±11.66	43.84 ±8.65	38.79 ±9.18	32.83 ±9.13	15.66 ±6.22	53.36±11.16	39.78±18.50	34.58
Qwen2.5-32B-I	19.49±11.17	7.42 ±6.16	20.08 ±9.81	35.85 ±9.25	29.82±11.19	39.50 ±8.48	32.63 ±8.40	24.55 ±8.20	13.87 ±6.75	45.12±12.65	38.14±15.37	27.86
Qwen-14B (EW-ours)	62.39±14.45	55.74±13.59	55.63±13.02	49.82 ±8.48	53.60±11.17	50.03 ±9.79	36.25 ±7.47	55.36±14.26	44.04 ±9.66	49.85 ±7.98	**66.27**±14.59	52.63
Qwen-32B (EW-ours)	**67.81**±12.50	**60.56**±12.20	61.34±10.76	**57.24** ±7.83	**58.62**±10.76	55.43 ±9.28	39.22 ±7.91	60.55±13.42	48.48 ±9.52	54.93 ±7.91	63.45±22.36	57.06
Qwen2.5-72B-I	25.71±10.42	9.55 ±5.89	30.88±10.99	37.77 ±8.08	38.90 ±6.64	46.31 ±7.83	36.60 ±7.27	36.50±14.50	23.44±10.29	43.86 ±9.17	41.56±32.31	33.73
Qwen3-32B	42.16±13.59	26.76±12.91	58.39±10.81	51.10 ±9.07	35.24±12.76	62.37±10.82	69.64±12.52	65.99 ±9.24	47.34±11.69	70.18 ±9.74	54.14±19.77	53.03
Llama-3.3-70B-I	22.66 ±8.93	6.23 ±4.10	27.03±11.61	36.36±10.95	40.42 ±5.32	46.90 ±7.07	39.27 ±8.92	39.83±11.62	22.21±12.14	52.85±11.61	46.33±28.49	34.55
Mistral-Small	54.19 ±9.61	39.24 ±9.77	57.23 ±7.62	46.02 ±5.48	44.53 ±8.14	68.46 ±6.27	70.89 ±7.01	61.67 ±6.76	40.48 ±6.78	52.40 ±9.10	63.85 ±5.01	54.45
DeepSeek-V3-0324	66.22±13.05	56.23±12.75	**75.59** ±8.32	55.73 ±7.73	53.28±11.00	**71.96** ±9.04	**76.43**±10.32	**72.48** ±8.38	**51.89**±14.65	**70.20** ±8.05	55.04±16.32	**64.10**

Table 2: Main benchmark results for **Character Agent** evaluation. I and P denote Instruct and Preview in model names. The best performances within the same model scale are **bold-faced**, and the second-best are underlined.

domain (OOD) test data from books excluded from training. Concatenating the extracted states along each timeline yields **138,596** supervised training samples for the seven tasks in §3.2; test samples are selected from specific time points, each containing the current character and world states from which simulation can continue. The test split contains **222** samples. Appendix D details dataset statistics.

3.4 Evaluation Framework

EvolvingWorld evaluates systems with two top-level score families, CHARACTER and WORLD, covering 10 dimensions and 20 metrics (on a 0–100 scale). These dimensions target persistent simulation quality beyond local persona imitation.

CHARACTER. The 6 dimensions are: (1) Character Consistency: Profile Fidelity, Speaking Style Fidelity, Motivation-Driven Behavior; (2) Evolution Quality: Profile Update Fidelity, Profile Evolution Smoothness; (3) Environmental Grounding: Environment Awareness, Environmental Utilization; (4) Interaction Quality: Contextual Responsiveness, Narrative Progression; (5) Motivation Generation: Motivation Quality; (6) Instruction Compliance: Instruction Compliance.

WORLD. The 4 dimensions are: (1) Scene Planning: Cast Selection Rationality, Location & Scenario Rationality, Scene Continuity & Coherence; (2) Speaker Management: Turn & Scene Orchestration; (3) World State Maintenance: Global Update Sensitivity, Global State Accuracy, Location Update Sensitivity, Location State Accuracy; (4) Instruction Compliance: Instruction Compliance.

For each trajectory τ, we use a **per-metric independent LLM-as-Judge** design, where each judge receives only relevant inputs and the rubric for one metric. Scores are aggregated hierarchically to capture both scene-level quality and long-range character/world evolution. If a simulation terminates early due to invalid output, we apply task-specific penalties so that the failures are reflected in both Instruction Compliance and the affected metrics. Appendix E gives the full scoring process.

4 Experiment

4.1 Experiment Setup

Training. We fine-tune open-source backbones with supervised instruction tuning. Following prior work (Wang et al., 2025b; Xu et al., 2026b), we mix in Tulu3 (Lambert et al., 2025) general instruction tuning data at a 1:1 ratio to preserve general capabilities. Our fine-tuned models span both model family and size, including *Llama-3.1-8B-Instruct*, *Qwen2.5-7/14/32B-Instruct*, and *Qwen3-4B-Instruct*. For brevity, we refer to each model by

SP: Scene Planning **SM**: Speaker Management **WSM**: World State Maintenance **IC**: Instruction Compliance
CSR: Cast Selection Rationality **LSR**: Location & Scenario Rationality **SCC**: Scene Continuity & Coherence
TSO: Turn & Scene Orchestration **GUS**: Global Update Sensitivity **GSA**: Global State Accuracy
LUS: Location Update Sensitivity **LSA**: Location State Accuracy

Models	SP			SM	WSM				IC	Avg.
	CSR	LSR	SCC	TSO	GUS	GSA	LUS	LSA	IC	
Closed-source										
Kimi-K2.5	78.36 ±7.78	80.61±15.01	51.72±25.95	58.84±10.95	66.64 ±7.53	**59.85** ±8.22	68.90 ±6.87	70.54 ±9.77	77.94 ±9.94	68.16
Gemini-2.5-Flash	73.18 ±6.95	84.09 ±6.10	66.36±14.98	43.02±24.56	60.41 ±7.49	56.78 ±4.10	54.22±11.27	56.25 ±6.31	43.49±23.83	59.76
Gemini-2.5-Pro	78.86 ±4.73	88.33 ±6.27	70.34±20.14	75.66 ±5.50	65.89 ±4.69	56.99 ±5.84	54.96 ±8.66	68.20 ±9.76	80.36 ±7.15	71.07
Gemini-3.1-Pro-P	83.50 ±3.17	91.60 ±4.81	69.16±20.97	72.70±10.18	65.60 ±9.89	55.13±10.68	66.51 ±9.67	61.54±10.84	85.55±13.49	72.37
GPT-4o	77.80 ±8.81	85.33 ±6.55	51.02±16.90	61.44 ±7.34	61.40 ±8.32	52.86 ±5.91	54.12 ±9.96	51.44 ±7.07	73.50±18.44	63.21
GPT-5-Chat	80.44 ±6.88	88.30 ±6.12	68.38±16.61	64.69±17.27	60.57 ±9.14	53.97 ±6.21	52.59 ±7.83	57.85 ±7.87	77.51±22.07	67.14
GPT-5.3-Chat	80.04 ±5.09	89.35 ±4.81	68.40±15.48	69.04 ±6.41	**70.02** ±2.12	51.11 ±5.48	**75.21** ±6.30	64.77±10.96	85.57 ±4.92	72.61
Claude-4.6-Sonnet	81.89 ±5.46	92.90 ±4.58	60.77±25.33	79.84±13.54	38.93±24.04	33.26±20.37	30.33±21.21	41.12±26.94	58.83±34.31	57.54
Claude-4.6-Opus	**84.86** ±5.34	**96.66** ±2.56	**82.28**±13.37	**84.98** ±4.76	66.05 ±5.74	58.48 ±6.75	61.72 ±9.46	**75.49** ±8.93	**89.34** ±6.50	**77.76**
Open-source										
< 14B										
Qwen3-4B-I	60.20±15.88	52.50±23.75	20.71±15.06	5.65 ±6.64	41.98±23.54	35.56±20.73	22.68±14.07	23.31±13.57	25.78±18.28	32.06
Qwen2.5-7B-I	61.07±12.78	51.25±14.71	14.09±14.59	5.82 ±4.92	53.53±11.52	44.53 ±9.69	28.65 ±9.43	31.53 ±8.02	35.58±16.83	36.23
Llama-3.1-8B-I	60.56±17.34	**68.01**±17.31	34.16±13.44	28.65±11.65	49.02±19.99	43.94±17.78	45.89±19.33	41.12±16.92	30.82±36.53	44.69
Qwen-4B (EW-ours)	61.78±10.20	62.42±11.14	23.20±16.20	30.81 ±6.65	57.62 ±7.97	44.15 ±8.01	58.09 ±7.59	46.55 ±6.40	72.24±13.19	50.76
Qwen-7B (EW-ours)	**63.90** ±8.89	63.32±11.66	32.02±17.41	**39.98** ±6.79	**61.21** ±5.41	**47.52** ±6.02	**61.20** ±4.10	**49.10** ±4.44	**76.59** ±9.78	**54.98**
Llama-8B (EW-ours)	63.58 ±7.91	67.57±11.51	**38.41**±23.29	39.79±11.31	56.59 ±7.62	42.89 ±7.78	60.19 ±5.80	46.75 ±4.99	72.90 ±7.92	54.30
≥ 14B										
Qwen2.5-14B-I	70.66±10.08	67.37±13.98	21.47±15.64	16.95 ±6.74	50.64±21.59	41.28±17.15	23.71±12.53	29.82±13.41	41.89±18.78	40.42
Qwen2.5-32B-I	69.55 ±9.80	60.18±14.46	17.69±14.67	18.03 ±7.14	61.21 ±6.80	49.53 ±7.00	42.35±13.03	42.08 ±7.58	51.12±17.99	45.75
Qwen-14B (EW-ours)	68.18 ±7.96	70.67±10.66	39.24±18.05	45.19±10.48	59.45 ±7.84	46.76 ±6.86	**60.84** ±6.83	49.08 ±6.37	**77.02**±15.15	57.38
Qwen-32B (EW-ours)	69.77 ±7.99	77.26 ±9.58	45.42±17.69	**48.21** ±9.83	61.67 ±3.38	**49.92** ±5.06	60.65 ±4.40	**51.95** ±3.96	73.94±24.58	**59.87**
Qwen2.5-72B-I	70.29 ±9.27	69.52 ±8.44	37.89±13.99	29.88±11.59	39.85±26.68	31.12±20.86	27.04±19.54	29.36±20.08	37.35±29.99	41.37
Qwen3-32B	67.59 ±9.82	71.76±12.88	34.70±20.55	43.60 ±9.59	54.41 ±8.87	47.33 ±7.26	53.52 ±9.16	49.02 ±9.48	51.34±20.40	52.59
Llama-3.3-70B-I	75.62±10.65	78.39±12.61	36.76±14.19	23.47 ±9.18	41.47±22.18	34.41±18.74	22.41±15.41	27.44±15.72	34.90±22.87	41.65
Mistral-Small	72.24 ±6.72	73.74 ±6.86	41.38±16.00	42.39 ±6.98	60.88 ±7.47	44.55 ±6.82	41.64 ±8.90	44.92 ±6.53	60.41 ±8.50	53.57
DeepSeek-V3-0324	**76.98** ±6.36	**81.86** ±8.43	**51.00**±19.30	38.75±15.05	**63.48** ±5.11	48.79 ±6.70	47.32±10.08	49.70 ±8.14	60.33±16.47	57.58

Table 3: Main benchmark results for **World Model** evaluation. I and P denote Instruct and Preview in model names. The best performances within the same model scale are **bold-faced**, and the second-best are underlined.

its size in the table. Details are in Appendix B.

Models. We evaluate 10 closed-source APIs, 11 open-source models, prior role-playing baselines including CoSER (Wang et al., 2025b) and Crab (He et al., 2025), and models trained on EvolvingWorld (EW). For untrained settings, the same LLM serves as Character Agent and World Model. For the trained ones (with their training data specified in parentheses), since CoSER and Crab only provide role-play data, we evaluate them as Character Agents paired with the same untrained backbone as World Model. EW models use separately trained Character Agent and World Model with the same backbone. Full results are in Appendix F.

Simulation. To evaluate long-horizon behavior, we allow up to 20 scenes per sample and 50 interactions per scene.

Appendix G reports ablation results, H reports in- and out-of-distribution results, I examines different judge-models, J demonstrates downstream video generation, and L reports human evaluation.

4.2 Main Results

Tables 2 and 3 report the mean and standard deviation scores. *Claude-4.6-Sonnet* is our main judge.

(1) EW gains come from modeling state evolution beyond role-play imitation. Across matched backbones, EW-training improves both Character Agent and World Model performance, showing the effect of book-to-world supervision. The advantage is especially clear over role-play-only baselines: on both Qwen-7B and Llama-8B, EW-trained Character Agents substantially outperform CoSER and Crab, indicating that long-horizon role-playing requires modeling coupled character and world-state evolution rather than only imitating dialogue.

(2) EW improves controlled character evolution and more precise environmental grounding. For Character Agents, the main gains concentrate on character consistency, evolution, and interactive progression. Models better preserve stable traits such as profile, speaking style, and motivations, while making profile updates and evolution smoother under event-conditioned changes. Although Environment Utilization sometimes decreases, Environment Awareness consistently improves, suggesting more accurate grounding rather than arbitrary use of environmental details.

(3) EW improves world modeling by targeting structured state tracking which is underrepresented in pretraining. For World Models, *Claude-4.6-Opus* achieves the best overall score, but models generally perform worse on world-model tasks than on character-agent tasks, reflecting the difficulty of explicit structured state tracking, a capability less covered by pretraining than narrative continuation. EW training improves this most clearly in long-range scene continuity, turn/scene orchestration, and location-level updates, showing that models learn to track how events reshape both global

CC: Character Consistency **EQ**: Evolution Quality **EG**: Environmental Grounding **IQ**: Interaction Quality **MG**: Motivation Generation
PF: Profile Fidelity **SSF**: Speaking Style Fidelity **MDB**: Motivation-Driven Behavior **PUF**: Profile Update Fidelity **PES**: Profile Evolution Smoothness
EA: Environment Awareness **EU**: Environmental Utilization **CR**: Contextual Responsiveness **NP**: Narrative Progression **MQ**: Motivation Quality

Models	Framework	CC			EQ		EG		IQ		MG	Avg.
		PF	SSF	MDB	PUF	PES	EA	EU	CR	NP	MQ	
Gemini-3.1-Pro-P	BookWorld	84.83 ±7.30	77.07 ±8.34	90.43 ±4.06	36.97 ±7.94	30.61 ±7.26	83.40 ±5.16	85.71 ±6.42	90.45 ±3.38	65.22±10.89	63.60±11.14	70.83
	EvolvingWorld	89.79 ±5.23	79.40 ±6.91	93.09 ±3.57	78.90 ±5.87	66.29 ±9.98	85.35 ±6.17	83.90 ±8.33	91.87 ±3.62	67.74 ±9.63	83.04 ±6.41	81.94
GPT-5.3-Chat	BookWorld	88.77 ±4.40	83.79 ±4.88	87.03 ±4.48	40.58 ±8.70	31.28±11.96	88.89 ±5.64	89.15 ±6.68	86.97 ±4.49	47.68 ±7.69	62.22±11.75	70.64
	EvolvingWorld	94.71 ±2.75	88.77 ±4.30	93.71 ±2.95	77.90 ±3.74	80.16 ±6.62	91.62 ±4.88	92.10 ±7.54	92.27 ±3.74	59.98 ±7.55	83.94 ±6.50	85.52
Llama-3.1-8B-I	BookWorld	7.19 ±6.67	4.55 ±4.14	9.30 ±5.86	7.86 ±4.57	10.08 ±6.66	23.09 ±8.94	34.27 ±8.46	13.92 ±6.87	10.58 ±7.21	12.99±12.22	13.38
	EvolvingWorld	30.58±11.10	22.30 ±8.36	30.71±12.65	21.87±16.04	20.36±13.89	44.33 ±8.88	46.19 ±9.30	40.78±12.36	40.77 ±9.04	27.26±22.46	32.52
Qwen2.5-14B-I	BookWorld	27.87±10.89	12.83 ±8.18	27.96 ±9.87	26.03 ±5.30	17.68 ±8.27	41.88 ±9.84	33.21 ±9.59	28.39 ±7.90	14.38 ±4.51	37.21 ±8.28	26.74
	EvolvingWorld	31.44±10.64	14.71 ±7.63	30.43±10.11	43.95 ±6.51	35.56±11.66	43.84 ±8.65	38.79 ±9.18	32.83 ±9.13	15.66 ±6.22	53.36±11.16	34.06

Table 4: Comparison of **Character Agent** performance between EvolvingWorld (ours) and BookWorld.

SP: Scene Planning **SM**: Speaker Management **WSM**: World State Maintenance **CSR**: Cast Selection Rationality
LSR: Location & Scenario Rationality **SCC**: Scene Continuity & Coherence **TSO**: Turn & Scene Orchestration
GUS: Global Update Sensitivity **GSA**: Global State Accuracy

Models	Framework	SP			SM	WSM		Avg.
		CSR	LSR	SCC	TSO	GUS	GSA	
Gemini-3.1-Pro-P	BookWorld	80.64 ±5.71	88.51 ±7.30	61.67±18.20	70.83 ±6.73	62.31 ±4.65	49.78 ±8.84	68.96
	EvolvingWorld	83.50 ±3.17	91.60 ±4.81	69.16±20.97	72.70±10.18	65.60 ±9.89	55.13±10.68	72.95
GPT-5.3-Chat	BookWorld	80.38 ±4.58	87.63 ±5.81	52.17±20.07	56.57 ±7.11	67.24 ±3.94	50.69 ±8.98	65.78
	EvolvingWorld	80.04 ±5.09	89.35 ±4.81	68.40±15.48	69.04 ±6.41	70.02 ±2.12	51.11 ±5.48	71.33
Llama-3.1-8B-I	BookWorld	42.21±11.95	44.16±13.19	4.49 ±8.17	5.09 ±2.85	45.72±10.74	30.46 ±8.08	28.69
	EvolvingWorld	60.56±17.34	68.01±17.31	34.16±13.44	28.65±11.65	49.02±19.99	43.94±17.78	47.39
Qwen2.5-14B-I	BookWorld	70.57 ±6.91	66.60 ±8.45	14.97±11.99	14.45 ±4.65	50.49 ±5.87	40.87 ±4.67	42.99
	EvolvingWorld	70.66±10.08	67.37±13.98	21.47±15.64	16.95 ±6.74	50.64±21.59	41.28±17.15	44.73

Table 5: Comparison of **World Model** performance between EvolvingWorld (ours) and BookWorld.

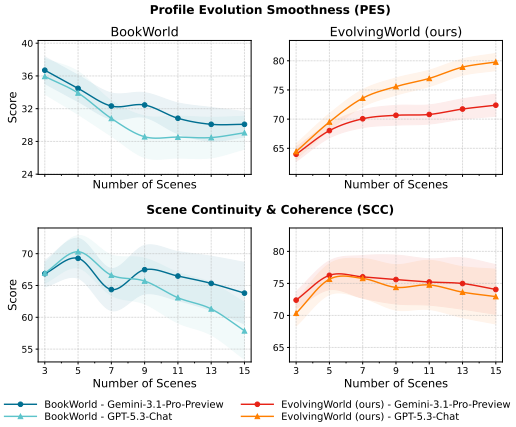


Figure 3: Length-wise comparison between **EvolvingWorld** and BookWorld on Profile Evolution Smoothness (PES) and Scene Continuity & Coherence (SCC). Shading denotes mean ± std/4.

and location-level states. With this targeted supervision, Qwen-32B trained on EW reaches a World Model Avg. of 59.87, surpassing even *Claude-4.6-Sonnet* and *Gemini-2.5-Flash*.

4.3 Comparison with BookWorld

We further compare EvolvingWorld with BookWorld (Ran et al., 2025), a representative long-horizon multi-agent role-play simulation baseline. BookWorld updates only a few predefined character fields, such as goals and states, while leaving broader profiles and relationships unchanged; its world agent also updates only a global event field and does not maintain world state changes. In contrast, we support open-schema character updates and world-state updates even to the entity level.

For a fair comparison, we evaluate both frameworks on our test set with four representative untrained backbones. Tables 4 and 5 report metrics shared by both frameworks. Across all backbones, EvolvingWorld achieves higher average scores for both character-agent and world-model evaluation, with especially large gains in character evolution metrics such as PUF, PES, and MQ. For world modeling, the main improvements appear in long-range scene continuity and turn/scene organization.

Figure 3 further examines PES and SCC as indicators of long-horizon evolution, showing that BookWorld degrades on both metrics as simulations become longer. But our EvolvingWorld mitigates this trend, and its PES even improves over longer trajectories, suggesting that structured state updates better support long-horizon evolution.

5 Conclusion

We presented EvolvingWorld, a framework and benchmark for simulating interactive, persistently evolving worlds. By coupling a Character Agent with a World Model, maintaining explicit open-schema states, and decomposing long-horizon interaction into seven trainable tasks, EvolvingWorld provides a concrete foundation for studying persistent world evolution beyond isolated role-play. Our results show that this design leads to more coherent long-horizon simulations across diverse backbones. We hope EvolvingWorld can serve as a useful step toward richer literary agents, controllable interactive worlds, and long-horizon role-playing systems.

Limitations

EvolvingWorld has three main limitations. First, it models the world as a single objective state shared by all characters, whereas literary characters often perceive and remember the same world differently. For example, one character may view the world as benevolent while another sees it as hostile, and a character may misremember an object as being on a table when it is actually on a chair. Such subjective perceptions and imperfect memories can shape character behavior, but maintaining separate perceived worlds for individual characters would substantially increase system complexity. We hope future work will address subjective world modeling in a comprehensive yet practical way.

Second, our world representation is constrained by the limited context length of LLMs. The World Model tracks only important entities at each location, rather than all entities in the environment. Future work may benefit from other modalities, such as information-dense visual inputs, to construct more detailed world representations.

Third, due to copyright constraints, our benchmark is built from public-domain classic books from Project Gutenberg, and does not cover domains such as modern novels, games, or user-created worlds. However, this is mainly a constraint of the data rather than of the framework. Since EvolvingWorld is fully open-schema and uses no domain-specific ontologies or predefined attribute slots, it can in principle adapt to other narrative domains. Even within the current corpus, the data already span multiple genres. And our OOD experiments (Appendix H) further show consistent gains on entirely unseen books. Broader cross-domain evaluation remains important future work.

Ethics Statement

Source Literary Works. EvolvingWorld is derived from public-domain literary works obtained from Project Gutenberg, a free digital library dedicated to digitizing and providing access to books whose copyrights have expired. The data consist of transformed and abstracted representations rather than verbatim reproductions of the original books.

Use of Human Annotations. Our institution recruited native-English-speaking annotators to assess the agreement between LLM judges and human judges. Throughout the annotation process, we ensured that annotators' privacy rights were respected and that their participation was voluntary and informed. Annotators received compensation exceeding the local minimum wage and consented to the use of the annotations they provided for research purposes. We also maintained transparency about the purpose of the study and how the annotations would be used. Appendix L provides further details on the human evaluation protocol and annotation instructions.

Risks. EvolvingWorld is derived from literary works and further augmented with LLM-generated annotations, both of which may raise ethical considerations. The source books may contain toxic, stereotypical, or discriminatory language, and LLMs may introduce or amplify biases during annotation. Therefore, although the data are used for research purposes, we cannot guarantee that they are entirely free from harmful content. The fictional content, character behaviors, and generated analyses in the dataset should not be interpreted as representing the authors' viewpoints. We encourage users of the dataset and benchmark to account for these risks and to avoid deploying the data or models in settings where harmful fictional content or biased generations could affect real individuals.

Data Use Restrictions. We emphasize that our dataset and benchmark are intended exclusively for scientific research and not for commercial use. We provide these resources solely for academic use and disclaim responsibility for any misuse beyond their intended research scope.

References

Jiaxin Bai, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, Feng Qin, Tianshi Zheng, Xi Peng, Xin Yao, Huiwen Yang, Leijie Wu, JI Yi, Gong Zhang, Renhai Chen, and Yangqiu Song. 2026. AutoSchemaKG: Autonomous knowledge graph construction through dynamic schema induction from web-scale corpora. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20557–20584, San Diego, California, United States. Association for Computational Linguistics.

Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. *Preprint*, arXiv:2505.13157.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset

for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 8506–8520. Association for Computational Linguistics.

Meng Chu, Xuan Billy Zhang, Kevin Qinghong Lin, Lingdong Kong, Jize Zhang, Teng Tu, Weijian Ma, Ziqi Huang, Senqiao Yang, Wei Huang, Yeying Jin, Zhefan Rao, Jinhui Ye, Xinyu Lin, Xichen Zhang, Qisheng Hu, Shuai Yang, Leyang Shen, Wei Chow, and 23 others. 2026. Agentic world modeling: Foundations, capabilities, laws, and beyond. *Preprint*, arXiv:2604.22748.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhenpeng Gao, Xiaofen Xing, and Xiangmin Xu. 2025. TailorRPA: A retrieval-based framework for eliciting personalized and coherent role-playing agents in general domain. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5381–5412, Suzhou, China. Association for Computational Linguistics.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A novel configurable role-playing LLM with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15030–15052. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *CoRR*, abs/2308.09597.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Xiaoteng Ma, Guanhua Chen, and Heng Ji. 2026. From word to world: Can large language models be implicit text-based world models? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8084–8111, San Diego, California, United States. Association for Computational Linguistics.

Jiayu Liu, Qihan Lin, Cheng Qian, Rui Wang, Emre Can Acikgoz, Xiaocheng Yang, Jiateng Liu, Zhenhailong Wang, Xiusi Chen, Heng Ji, and 1 others. 2026a. Planbench-xl: Evaluating long-horizon planning of llm tool-use agents in large-scale tool ecosystems. *arXiv preprint arXiv:2606.22388*.

Jiayu Liu, Cheng Qian, Zhaochen Su, Qing Zong, Shijue Huang, Bingxiang He, and Yi R. Fung. 2026b. CostBench: Evaluating multi-turn cost-optimal planning and adaptation in dynamic environments for LLM tool-use agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12826–12858, San Diego, California, United States. Association for Computational Linguistics.

Jiayu Liu, Cheng Qian, Zhenhailong Wang, Bingxuan Li, Jiateng Liu, Heng Wang, Jeonghwan Kim, Yumeng Wang, Xiusi Chen, Yi R Fung, and 1 others. 2026c. Adaplanbench: Evaluating adaptive planning in large language model agents under world and user constraints. *arXiv preprint arXiv:2606.05622*.

Jiayu Liu, Rui Wang, Qing Zong, Qingcheng Zeng, Tianshi Zheng, Haochen Shi, Dadi Guo, Baixuan Xu, Chunyang Li, and Yangqiu Song. 2026d. Naacl: Noise-aware verbal confidence calibration for llms in rag systems. *arXiv preprint arXiv:2601.11004*.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–221, Vienna, Austria. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7828–7840. Association for Computational Linguistics.

Yuxing Lu, Wei Wu, Xukai Zhao, Rui Peng, and Jinzhuo Wang. 2026. Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. *Preprint*, arXiv:2502.06472.

Joon Sung Park, Joseph C. 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, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2026. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. BOOKWORLD: from novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15898–15912. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13153–13187. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11836–11850. Association for Computational Linguistics.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6372–6391. Association for Computational Linguistics.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 14743–14777. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025b. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Zixiao Wang, Duzhen Zhang, Ishita Agarwal, Shen Gao, Le Song, and Xiuying Chen. 2025c. Beyond profile: From surface-level facts to deep persona simulation in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21239–21257, Vienna, Austria. Association for Computational Linguistics.

Zhenhua Xu, Dongsheng Chen, Jian Li, Yitong Lin, Zhebo Wang, Jiafu Wu, Yizhang Jin, Chengjie Wang, Meng Han, and Yabiao Wang. 2026a. Improving general role-playing agents via psychology-grounded reasoning and role-aware policy optimization. *arXiv preprint arXiv:2606.27025*.

Zhenhua Xu, Dongsheng Chen, Shuo Wang, Jian Li, Chengjie Wang, Meng Han, and Yabiao Wang. 2026b. Adamarp: An adaptive multi-agent interaction framework for general immersive role-playing. *CoRR*, abs/2601.11007.

Ming Yan, Ruihao Li, Hao Zhang, Hao Wang, Zhilan Yang, and Ji Yan. 2023. LARP: language-agent role play for open-world games. *CoRR*, abs/2312.17653.

Bohao Yang, Dong Liu, Chenghao Xiao, Kun Zhao, Chen Tang, Chao Li, Lin Yuan, Yang Guang, and Chenghua Lin. 2025. Crafting customisable characters with llms: A persona-driven role-playing agent framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20216–20240. Association for Computational Linguistics.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025a. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11992–12022. Association for Computational Linguistics.

Yichen Yu, Yifan Jiang, Mandy Lui, and Qiao Jin. 2025b. Genlarp: Enabling immersive live action role-play through llm-generated worlds and characters. *Preprint*, arXiv:2510.14277.

Zeyu Zhang, Jianxun Lian, Chen Ma, Yaning Qu, Ye Luo, Lei Wang, Rui Li, Xu Chen, Yankai Lin, Le Wu, Xing Xie, and Ji-Rong Wen. 2025. TrendSim: Simulating trending topics in social media under poisoning attacks with LLM-based multi-agent

system. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2930–2949, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 1457–1476. Association for Computational Linguistics.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Characterbench: Benchmarking character customization of large language models. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26101–26110. AAAI Press.

Qing Zong, Jiayu Liu, Tianshi Zheng, Chunyang Li, Baixuan Xu, Haochen Shi, Weiqi Wang, Zhaowei Wang, Chunkit Chan, and Yangqiu Song. 2025a. Critical: Can critique help llm uncertainty or confidence calibration? *arXiv preprint arXiv:2510.24505*.

Qing Zong, Zhaowei Wang, Tianshi Zheng, Xiyu Ren, and Yangqiu Song. 2025b. ComparisonQA: Evaluating factuality robustness of LLMs through knowledge frequency control and uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4101–4117, Vienna, Austria. Association for Computational Linguistics.

Appendices

A Partial Support Clarifications

This section provides the rationale for the partial-support marks in Table 1.

BookWorld (Ran et al., 2025). We mark BookWorld as partial support for profile updates because it updates only several character fields, while leaving broader profile information and relations unchanged. For global world state updates, BookWorld maintains a global event, but other components of the world setting remain static. For location/entity-level state modeling, BookWorld provides only a brief description for each location, without entity-level modeling or updates to these world states.

GenerativeAgents (Park et al., 2023). We mark GenerativeAgents as partial support for profile updates because it updates a character's memory and then queries the memory to infer the character's latest status, so the updated information is still limited to predefined fields. We also mark it as partial support for scene initialization because its simulation proceeds at the granularity of days: all characters act each day, which differs from interactive literary worlds where updates occur scene by scene and each scene usually involves only a subset of characters. Its updates are therefore closer to everyday routines and minor daily activities than to scene-level literary progression.

CharacterBox (Wang et al., 2025a). We mark CharacterBox as partial support for profile updates because it mainly updates predefined attributes such as BDI (Belief-Desire-Intention) and position, while leaving broader Profile & Traits unchanged. We also mark it as partial support for location/entity-level state modeling because it simulates a single scene, with only one location-level environment description and state update, and therefore does not support long-horizon state updates across multiple locations and scenes.

B Training and Inference Details

We fine-tune all EW models with supervised fine-tuning using LoRA in LLaMA-Factory[2]. Unless otherwise specified, we use LoRA rank 64, LoRA alpha 128, and LoRA dropout 0.05. We train for 2 epochs with a learning rate of 2×10^{-5}, a maximum sequence length of 32,768 tokens, a per-

[2] https://llamafactory.readthedocs.io/en/latest/

device batch size of 1, and gradient accumulation of 64, yielding an effective batch size of 64 per GPU. We use bf16 precision, gradient checkpointing, and FlashAttention-2, and hold out 10% of the merged training data as the validation set. After training, we serve models with vLLM[3] for test-time inference and use 50 concurrent requests.

C Data Construction Details

This appendix expands the data construction pipeline summarized in §3.3.1. We first present the three main construction phases, then describe additional cleaning steps that are interleaved in implementation but separated here for readability.

C.1 Main Construction Phases

Book Selection. Following Coser (Wang et al., 2025b) and AdaMARP (Xu et al., 2026b), we select 57 representative books from Goodreads' Best Books Ever list[4] and obtain their full texts from Project Gutenberg[5], a free digital library of public-domain books. All selected books are narrated in chronological order, which ensures that the temporal progression of scenes aligns with the reading order, a prerequisite for our scene-by-scene state-tracking pipeline.

Phase 1: Scene Extraction. The book text is split into chunks for LLM processing. Each chunk is used to extract structured scenes, where each scene contains a summary, a scenario description, a list of key characters with brief descriptions, and a sequence of multi-turn interactions. Each interaction is represented as an actor-content pair: the actor can be a single character or a character group, allowing events such as an entire family entering together to be captured as one shared interaction. The content may include thought, speech, and action, marked consistently with the simulation format: thoughts are enclosed in `[...]`, speech is written as plain text, and actions are enclosed in `(...)`. The extraction supports cross-chunk continuation: when a scene is truncated at the end of one chunk, the truncated text is prepended to the next chunk, so the LLM can continue the same scene with the missing context.

Phase 2: Character Construction. For each character, we provide the LLM with extracted mentions and descriptions to identify aliases and determine a single official name (e.g., "Mr. Smith", "John", "Father" → "John Smith"). All references in scenes are then replaced with standardized names for cross-scene consistency. To initialize characters, we provide the LLM with the first few scenes involving each character and ask it to summarize an open-schema profile before the main story events unfold. Finally, profiles are updated scene by scene: after each scene, the LLM revises the profile, short description, and hidden tracker for every participating character based on what occurred and what later scenes reveal.

Phase 3: World Construction. Locations are extracted from scenes and undergo the same alias-standardization process as characters. The pipeline then generates an initial world state with a global state and per-location states. World states are updated interaction by interaction: after each interaction within a scene, the LLM decides whether persistent changes to the global or location state should be recorded.

Statistic	Count
Extracted Structured Data	
Total Books	57
Total Scenes	9,763
Total Interactions	132,800
Extracted Entities	
Unique Characters	3,311
Unique Locations	1,888
Constructed Dataset	
Train	138,596
Test (ID / OOD)	222 (116 / 106)

Table 6: Overview of EvolvingWorld dataset statistics.

C.2 Interleaved Cleaning Details

The following steps are shown after the main phases to keep the high-level pipeline clear, but they are interleaved in implementation. In particular, scene and interaction cleaning is applied after initial scene extraction and before the cleaned scenes are used for character construction, world construction, and task generation.

Interaction Refinement. Each interaction's content is sent to an LLM for refinement. The goal is to improve the clarity, consistency, and readability of interaction expressions while preserving their semantic meaning and narrative voice. In particular, the refinement removes redundant thoughts that are accidentally embedded in speech, keeps private thoughts separated from spoken utterances, and normalizes each interaction into the first-person perspective of its acting character or character group.

[3] https://vllm.ai/

[4] https://www.goodreads.com/list/show/1.Best_Books_Ever

[5] https://www.gutenberg.org/

Selected Books		
1. *A Doll's House*	**2.** *A Little Princess*	**3.** *A Tale of Two Cities*
4. *Anthem*	**5.** *Black Beauty*	**6.** *Don Quixote*
7. *Dr Jekyll and Mr Hyde*	**8.** *Far From the Madding Crowd*	**9.** *Great Expectations*
10. *Gulliver's Travels*	**11.** *Heart of Darkness*	**12.** *Jude the Obscure*
13. *Julius Caesar*	**14.** *Little Women*	**15.** *Madame Bovary*
16. *Mansfield Park*	**17.** *Middlemarch*	**18.** *Much Ado About Nothing*
19. *My Ántonia*	**20.** *Northanger Abbey*	**21.** *Notes from Underground*
22. *Oliver Twist*	**23.** *Othello*	**24.** *Pride and Prejudice*
25. *Sense and Sensibility*	**26.** *Siddhartha*	**27.** *Tess of the D'Urbervilles*
28. *The Adventures of Tom Sawyer*	**29.** *The Age of Innocence*	**30.** *The Call of the Wild*
31. *The House of Mirth*	**32.** *The Jungle*	**33.** *The Metamorphosis*
34. *The Phantom of the Opera*	**35.** *The Picture of Dorian Gray*	**36.** *The Pilgrim's Progress*
37. *The Portrait of a Lady*	**38.** *The Scarlet Letter*	**39.** *The Secret Garden*
40. *The Sorrows of Young Werther*	**41.** *The Sun Also Rises*	**42.** *The Tempest*
43. *The Three Musketeers*	**44.** *The Turn of the Screw*	**45.** *The Wind in the Willows*
46. *Treasure Island*	**47.** *Uncle Tom's Cabin*	**48.** *White Fang*
49. *A Portrait of the Artist as a Young Man*		
50. *Alice's Adventures in Wonderland / Through the Looking-Glass*		
51. *Around the World in Eighty Days*		
52. *The Adventures of Huckleberry Finn*		
53. *The Adventures of Sherlock Holmes (Sherlock Holmes, #3)*		
54. *The Hound of the Baskervilles (Sherlock Holmes, #5)*		
55. *The Importance of Being Earnest*		
56. *The Murder of Roger Ackroyd (Hercule Poirot, #4)*		
57. *Twenty Thousand Leagues Under the Sea (Captain Nemo, #2)*		

Table 7: The 57 selected books from Goodreads' Best Books Ever list.

Consecutive interactions performed by the same set of characters are also merged into a single interaction, removing artificial turn boundaries introduced by the chunked extraction process and producing more natural multi-turn sequences.

Duplicate Scene Removal. Because chunk boundaries can cause the same narrative event to be extracted twice (once at the end of one chunk and once at the beginning of the next), an LLM is used to detect pairs of scenes describing the same event. For each duplicate pair, the scene with less information is removed while the richer version is retained.

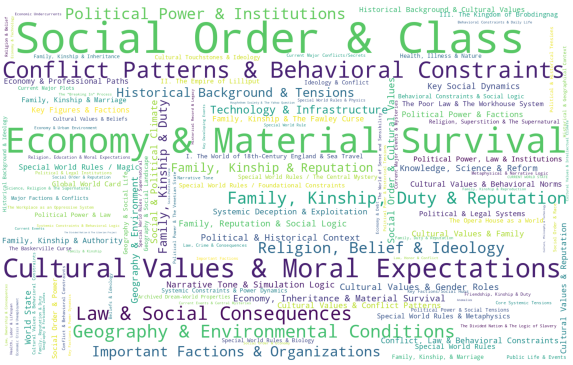
Scene Enhancement. Each scene's scenario description is enhanced by an LLM to add dramatic setup details, atmospheric context, and implicit tensions that were present in the source text but not captured during initial extraction. This produces richer scenario descriptions for the `location_scenario` training task.

C.3 Training and Test Data Construction

After obtaining the character states, global world state, and location states sequentially updated along the book timeline, we convert the timeline into task-level supervision according to the simulator's execution order. At a scene boundary, `scene_cast` takes the previous scene content together with the current global state, location states, and character states, and predicts whether the next scene exists and which characters participate; conditioned on this cast, `location_scenario` predicts the next location and scenario. Once the next scene is fixed, `motivation_update` uses the previous scene, the planned next scene, and the character state to infer each participating character's entering motivation.

Within a scene, each interaction history is used as input to supervise the next actor selection and interaction generation. `next_character` predicts the next actor or scene termination from the current scenario and visible interaction history. `interaction_gen` then generates that actor's next interaction from the same history together with the actor's character state and motivation. After each generated interaction, `world_update` decides whether the latest interaction requires a persistent update to the global or location state. Finally, after a scene is completed, `character_update` updates each relevant character state from the completed scene. Since `interaction_gen` and `character_update` are naturally formulated as multi-turn conversations, we store them as ShareGPT-style conversation examples following CoSER (Wang et al., 2025b) and AdaMARP (Xu et al., 2026b); the remaining tasks use task-specific prompts with structured targets under the same SFT interface. In this way, one extracted timeline is decomposed into seven super-

(a) Character Profile Dimensions



(b) Global World State Dimensions

Figure 4: Word clouds of state dimensions under the open-schema design for character profiles (left) and global world states (right). Font size is proportional to frequency across the 57-book corpus.

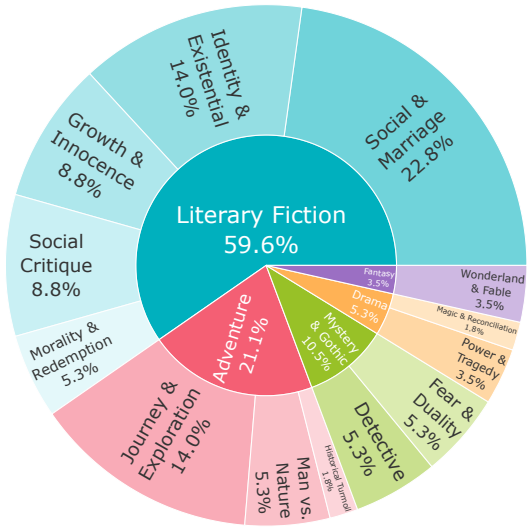


Figure 5: Two-level genre distribution of the 57 books in EvolvingWorld. The inner ring shows five coarse genres; the outer ring breaks each genre into thematic sub-categories.

vised tasks defined in §3.2.

For evaluation, we instead extract simulation snapshots at selected time points. Each snapshot stores the current character states, global state, and location states, together with the previous scene used for reference. We split books into three groups: 10% are held out as OOD books and never used for training; among the remaining books, half are train/test books and half are train-only books. For train/test books, the first 70% of scenes are used for training and snapshots are sampled from the remaining 30% as in-domain tests. Train-only books contribute all scenes to training. OOD books contribute no training examples; their test snapshots are sampled from the latter 70% of scenes. We sample only valid candidate scenes with at least five interactions, using at most five ID snapshots per train/test book and at most twenty OOD snapshots per OOD book.

D Dataset Statistics

D.1 Overview

Table 6 summarizes the key statistics of EvolvingWorld dataset. Table 7 lists all 57 selected books.

Schema Distribution. A distinctive feature of EvolvingWorld is that both character profiles and global world states adopt an *open schema*: rather than prescribing a fixed set of attributes, the extraction model freely selects whichever dimensions are most salient for a given character or narrative context. Figure 4 visualizes the resulting dimension vocabularies as word clouds, where font size reflects frequency across the corpus. Character profiles (left) are dominated by dimensions such as *Social Standing*, *Core Personality*, *Key Relationships*, and *Professional Identity*, while world states (right) center on *Cultural Values & Moral Expectations*, *Social Order & Class*, and *Economy & Material Survival*. The long tail of less frequent dimensions (581 unique character dimensions, 136 world-state dimensions) demonstrates the schema's flexibility in capturing diverse narrative elements.

Genre Distribution. To characterize the literary diversity of the corpus, we categorize all 57 books along two levels: a coarse *genre* layer and a finer *thematic* layer. As shown in Figure 5, the collection is dominated by Literary Fiction (59.6%), followed by Adventure (21.1%), Mystery & Gothic (10.5%), Drama (5.3%), and Fantasy (3.5%). Within each genre, thematic sub-categories further distinguish works by narrative focus. For example, Literary

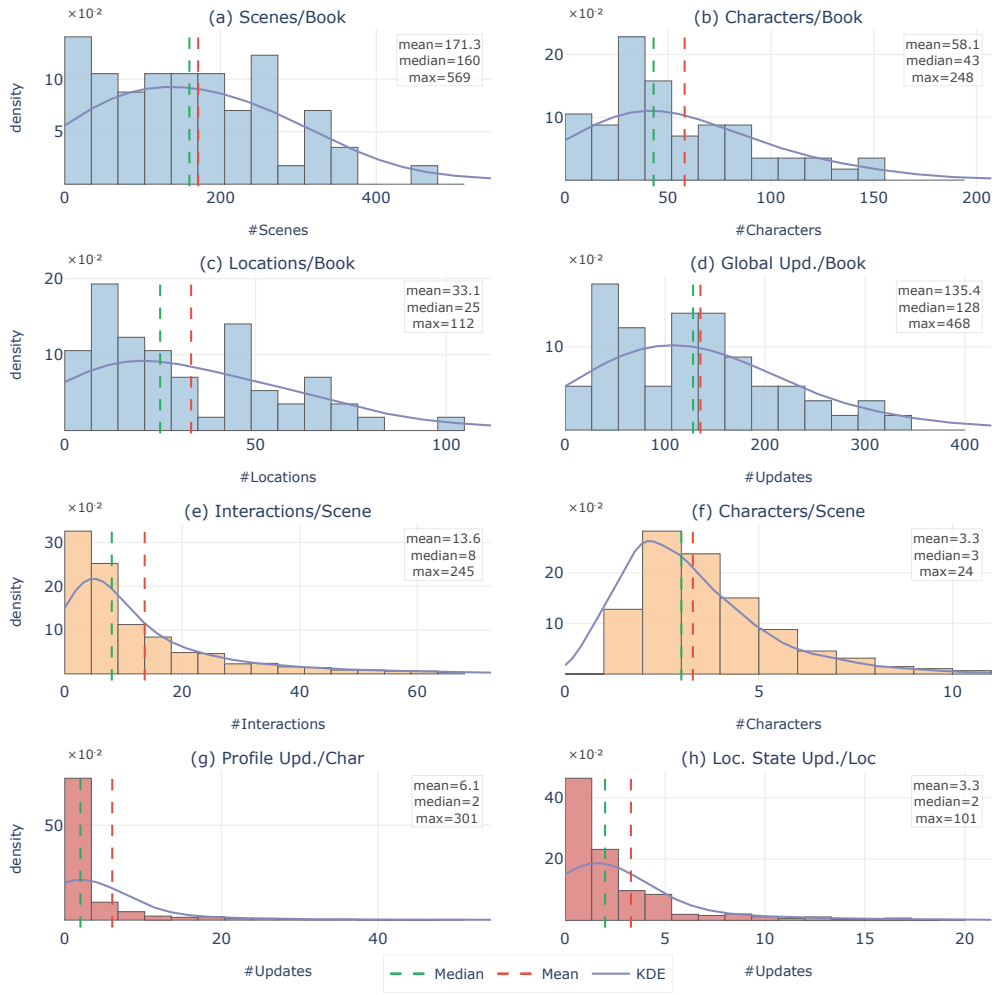


Figure 6: Distribution of extracted data statistics across four granularity levels. Each panel shows a histogram with KDE curve; dashed lines indicate the mean (red) and median (blue).

fiction spans Social & Marriage, Identity & Existential, Growth & Innocence, Social Critique, and Morality & Redemption. This breadth ensures that the benchmark exercises a wide range of character dynamics, world-building patterns, and narrative structures.

Extracted Data Distribution. Figure 6 presents the distribution of eight key statistics across the extracted structured data, organized by four granularity levels. At the *book level*, the number of scenes, characters, and locations per book varies considerably (median 160 scenes, 43 characters, and 25 locations), reflecting the diverse scale and complexity of the source novels. At the *scene level*, each scene typically involves a small number of interactions and characters, though long-tail cases exist for densely populated scenes. At the *character level*, profile update counts capture how frequently each character evolves throughout the narrative. At the *world level*, global state updates per book and location-specific state updates per location quantify the dynamism of the story world. Together, these distributions confirm that EvolvingWorld covers a broad spectrum of narrative complexity, from compact novellas to sprawling multi-character epics.

D.2 Training Data Statistics

Table 8 summarizes the training data for each task of the World Model and Character Agent. The "Samples" column reports the number of ShareGPT-format training conversations. For multi-turn tasks (`next_character` and `interaction_gen`), each sample contains multiple user–assistant exchanges within a single conversation; the "Asst. Turns" column gives the total number of assistant responses across all samples. Single-turn tasks have exactly one assistant turn per sample.

Token Length Distribution. Figure 7 shows the per-sample token length distribution for each of the

Module / Task	Samples	Asst. Turns
World Model		
`scene_cast`	7,983	7,983
`location_scenario`	7,958	7,958
`next_character`	14,315	116,789
`world_update`	17,832	17,832
Subtotal	48,088	150,562
Character Agent		
`interaction_gen`	40,554	108,831
`character_update`	24,977	24,977
`motivation_update`	24,977	24,977
Subtotal	90,508	158,785
Total	**138,596**	**309,347**

Table 8: Training data statistics per task. "Samples" = number of ShareGPT style conversations; "Asst. Turns" = total assistant responses (equals Samples for single-turn tasks).

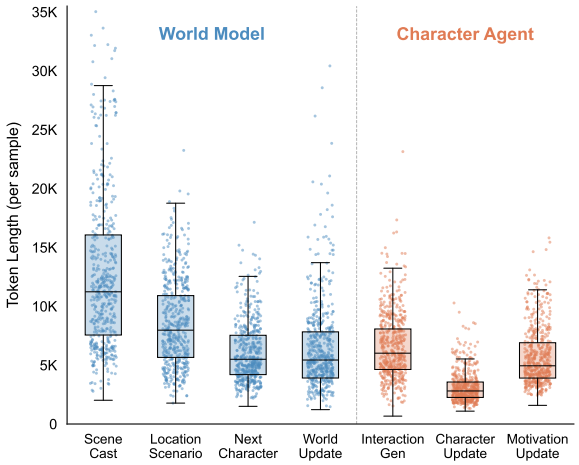


Figure 7: Token length distribution per task.

seven tasks, measured with the `cl100k_base` tokenizer (GPT-4 family). Each sample corresponds to a complete ShareGPT-format conversation, including all system, user, and assistant turns.

`scene_cast` exhibits the longest sequences (median $\approx$ 11.2k) because it must select the next scene's cast from *all* characters; to control input length, each character is represented by the latest short description rather than the full profile. `location_scenario` follows (median $\approx$ 8.0k) for an analogous reason: it chooses the scene location from *all* candidate locations; here each candidate location is represented by its description, without expanding the states of all contained entities. The four mid-range tasks, `next_character`, `world_update`, `interaction_gen`, and `motivation_update` (medians 4.9k–6.0k), are more compact because none of them requires enumerating all characters or all locations simultaneously. `character_update` is the shortest (median $\approx$ 2.8k) as it only updates a single character's profile based on one scene's events.

These distributions inform our choice of maximum sequence length and packing strategy for supervised fine-tuning. We accordingly set the training cutoff length to 32,768 tokens to accommodate the long-tail samples while keeping computation tractable.

E Evaluation Framework

E.1 Simulation Protocol

We evaluate EvolvingWorld with a **multi-scene simulation protocol** designed for persistent book-to-world simulation. Starting from each held-out structured snapshot, we run the simulator forward to generate a multi-scene trajectory containing scene plans, interaction histories, world-state updates, and character-state updates. This design allows us to evaluate not only local response quality, but also whether a model can maintain coherent world and character evolution across scenes.

Concretely, each evaluation sample provides the initial simulator state, including the current World State, character Full Profiles, and source scene. The evaluated system then produces a complete generated trajectory which includes the scene cast and location-scenario plan for each scene, the within-scene interaction sequence, the updated World State, and the updated Full Profiles written back after the scene ends.

Algorithm 1 summarizes the full simulation procedure.

E.2 Trajectory-Level Evaluation

Because EvolvingWorld introduces explicit persistent states and structured simulator outputs, scene-level dialogue evaluation (Xu et al., 2026b) alone is insufficient. We therefore adopt a **trajectory-level LLM-as-Judge protocol** that scores both the **Character Agent** and the **World Model** on complete simulated trajectories containing many scenes rather than isolated responses. The evaluation system comprises **10 dimensions and 20 sub-metrics**, divided into two scoring modules:

- **CHARACTER Score** (Character Agent Evaluation): 6 dimensions, 11 sub-metrics
- **WORLD Score** (World Model Evaluation): 4 dimensions, 9 sub-metrics

Algorithm 1 EvolvingWorld simulation with a World Model and a Character Agent

Require: Initial states $S_w^{\mathcal{L},(0)}$, $S_c^{\mathcal{I},(0)}$; World Model policies π_w; Character Agent policies π_c; max scenes T; max turns K

$\tau \leftarrow [\,]$
for $t = 1$ to T **do**
  $z_t \sim \pi_{\text{cast}}(\cdot \mid O_w^{\mathcal{I},\varnothing,(t)})$
  if $z_t = \varnothing$ **then**
    break
  end if
  $r_t \sim \pi_{\text{loc}}(\cdot \mid O_w^{z_t,\mathcal{L},(t)}, z_t)$; let ℓ_t be its location
  for all $i \in z_t$ **do**
    $M^{i,(t)} \sim \pi_{\text{mot}}(\cdot \mid O_c^{i,\ell_t,(t)}, r_t)$
  end for
  $Y_t \leftarrow [\,]$, $S_w^{\ell_t,(t,1)} \leftarrow S_w^{\ell_t,(t)}$
  for $k = 1$ to K **do**
    $i_{t,k} \sim \pi_{\text{next}}(\cdot \mid O_w^{z_t,\{\ell_t\},(t,k)}, r_t, Y_t)$
    if $i_{t,k} =$ `END` **then**
      break
    end if
    $y_{t,k} \sim \pi_{\text{int}}(\cdot \mid O_c^{i_{t,k},\ell_t,(t)}, r_t, Y_t)$
    $Y_t \leftarrow Y_t \,\|\, [y_{t,k}]$
    $S_w^{\ell_t,(t,k+1)} \sim \pi_{\text{wu}}(\cdot \mid O_w^{z_t,\{\ell_t\},(t,k)}, y_{t,k}, Y_t)$
  end for
  $K_t \leftarrow |Y_t|$, $S_w^{\ell_t,(t+1)} \leftarrow S_w^{\ell_t,(t,K_t+1)}$
  for all $i \in z_t$ **do**
    $S_c^{i,(t+1)} \sim \pi_{\text{cu}}(\cdot \mid O_c^{i,\ell_t,(t)}, Y_t, S_w^{\ell_t,(t+1)})$
  end for
  Carry forward unchanged states for $i \notin z_t$ and $\ell \neq \ell_t$
  Append $(z_t, r_t, Y_t, O_w^{\mathcal{I},\mathcal{L},(t+1)})$ to τ
end for
return τ

For **Character Agent** evaluation, the judge is given the relevant character profile information, scene motivations, world state context, and the generated trajectory segments needed for that metric. For **World Model** evaluation, the judge receives the initial simulator context together with the generated scene plans, turn orchestration decisions, and persistent state updates. This separation mirrors the modular structure of EvolvingWorld and allows us to diagnose planning errors and embodiment errors independently.

Each sub-metric is scored on a 0–100 scale with a base score of 50. The judge first identifies **Merits** (excellent aspects, each awarded 1 to 10) and then **Demerits** (problematic aspects, each penalized 1 to 10). The final score is computed as $\min(100, \max(0, 50 + \sum \text{merits} - \sum \text{demerits}))$.

Error Penalties. When a simulation terminates prematurely because a model fails to follow output format specifications, we apply two complementary penalties to ensure the error is reflected in the final scores. The error message contains the failing task name (e.g., `world_update`, `interaction_gen`), which we use to attribute the fault to either the World Model or the Character Agent. Infrastructure errors (e.g., API/Network error) are never attributed to the model and incur no penalty.

Failing Task	Penalized Metrics
`scene_cast`	CSR, SCC
`location_scenario`	LSR, SCC
`next_character`	TSO
`world_update`	GUS, GSA, LUS, LSA
`interaction_gen`	PF, SSF, MDB, EA, EU, CR, NP
`character_update`	PUF, PES
`motion_update`	MQ

Table 9: Task-to-metric mapping for the Metric Penalty. When a task causes simulation termination, its corresponding metrics receive the additional penalty described above.

(1) IC Penalty (Instruction Compliance). We penalize the IC score of the responsible model using a logarithmic-decay formula based on n, the total number of calls made by that model before the failure point:

$$\text{penalty}_{\text{IC}} = \min\left(50,\ \frac{50}{\ln(n+1)}\right) \tag{1}$$

The final IC score is $\max(0,\ \text{IC}_{\text{judge}} - \text{penalty}_{\text{IC}})$, where IC_{judge} denotes the raw IC score assigned by the LLM judge before any penalty. The logarithmic decay provides a smooth curve: a model that fails on its very first call receives the maximum penalty of 50 (driving IC to zero), while a model that succeeds hundreds of times before failing still incurs a meaningful penalty of ∼8–10 points. This ensures that "the model eventually crashed due to format non-compliance" is always reflected in the score, regardless of how many successful calls preceded the failure.

(2) Metric Penalty (Task-Specific Metrics). Beyond IC, the failing task also maps to specific quality metrics via a fixed task-to-metric correspondence (Table 9). For these metrics, we apply a penalty that simulates the effect of an additional failed scene:

$$\text{score} = \text{score}_{\text{judge}} \times \frac{N}{N+1} \tag{2}$$

where $\text{score}_{\text{judge}}$ is the average score assigned by the LLM judge over the N successfully completed scenes, and score is the final reported score for that metric. This is equivalent to appending a zero-score virtual scene to the existing average. For samples that fail before producing any scene ($N = 0$), the affected metrics are set directly to 0. IC metrics

	Character Agent (CHARACTER)		World Model (WORLD)
Dim I	Character Consistency (PF, SSF, MDB)	**Dim I**	Scene Planning (CSR, LSR, SCC)
Dim II	Evolution Quality* (PUF, PES)	**Dim II**	Speaker Management (TSO)
Dim III	Environmental Grounding (EA, EU)	**Dim III**	World State Maintenance* (GUS, GSA, LUS, LSA)
Dim IV	Interaction Quality (CR, NP)	**Dim IV**	Instruction Compliance (IC)
Dim V	Motivation Generation (MQ)		
Dim VI	Instruction Compliance (IC)		
Total	**11 sub-metrics**		**9 sub-metrics**

Table 10: Overview of the evaluation framework. * denotes evaluation perspectives introduced by this framework. Grand total: 10 dimensions, 20 sub-metrics.

are excluded from this penalty (already handled above).

At aggregation time, scene-local metrics are scored for each generated scene and averaged across relevant scenes. Character-level trajectory metrics are aggregated over the scenes in which a character participates; specifically, **Profile Evolution Smoothness** is scored over each character's participating scenes and then averaged with weights proportional to character appearances. Full-trajectory metrics are computed over the complete generated trajectory; specifically, **Scene Continuity & Coherence** is scored over the full simulation to measure long-range scene organization and world consistency. The final **CHARACTER** and **WORLD** scores are simple averages over their corresponding sub-metrics, with missing metrics treated as invalid rather than silently imputed.

E.3 Evaluation Overview

Table 10 provides a summary of all evaluation dimensions and sub-metrics.

E.4 Character Agent Evaluation (CHARACTER Score)

E.4.1 Dim I: Character Consistency

Evaluates whether the character consistently embodies its profile throughout the interaction. Core question: if character's name was hidden, would the output still be recognizable as the character?

Profile Fidelity (PF). Whether the character's knowledge, skills, and behavior are strictly confined within the profile scope. Criteria: (a) *Knowledge Boundaries*: no knowledge or skills beyond profile settings (e.g., an ordinary farmer should not demonstrate advanced magical theory); (b) *Background Consistency*: behavior matches the character's age, social class, and historical background; (c) *Ability Constraints*: no abilities or privileges not documented in the profile appear out of nowhere.

Speaking Style Fidelity (SSF). Whether the character's speaking style is consistent with the profile, natural, and free of obvious AI artifacts. Criteria: (a) *Style Markers*: language features defined in the profile are used (catchphrases, sentence patterns, terminology, dialects); (b) *Emotional Tone*: tone matches the character's personality rather than a generic "AI assistant" tone; (c) *Naturalness*: language is fluent, human-like, avoiding templated or mechanical artifacts.

Motivation-Driven Behavior (MDB). Whether the core motivation continuously drives decisions, and whether thought/action/speech are logically consistent. Criteria: (a) *Behavioral Attribution*: decisions traceable to core motivation rather than random unmotivated behavior; (b) *Trinity Coherence*: thought reasonably drives action and speech, with no unexplained contradictions among the three; (c) *Value Stability*: core values do not suddenly change without a significant plot trigger.

E.4.2 Dim II: Evolution Quality*

Evaluates whether the dynamic updates to the character profile and hidden tracker are reasonable, coherent, and complete. Unique to this framework, focusing on character growth quality during interactions.

Profile Update Fidelity (PUF). Whether the profile update and hidden tracker update jointly capture all changes that should be preserved from the scene, with appropriate threshold judgment between the two. Criteria: (a) *Causal Chain*: every item written to profile or hidden tracker has a clear triggering event in the scene as evidence, with no fabricated information (Liu et al., 2026d); (b) *Growth/Signal Capture*: important changes that have crossed the threshold are written to profile; subtle signals that may accumulate in the future are written to hidden tracker; (c) *Threshold Judgment*: major persistent changes should not remain only in the tracker without updating the profile; slight or ambiguous signals should not be prematurely written into the profile; (d) *No Over/Under-Updating*:

profile remains concise and stable, hidden tracker does not become a meaningless log; together they neither miss important changes nor over-react to irrelevant details.

Profile Evolution Smoothness (PES). Whether the profile and hidden tracker jointly exhibit gradual, coherent, and appropriately-scaled evolution across scenes. Criteria: (a) *Magnitude Matching*: casual conversations should only leave light signals or no change; major events trigger significant profile updates; sub-threshold content settles in the tracker first; (b) *Gradualness*: personality, attitude, and relationship changes go through reasonable transitional stages; hidden tracker preserves intermediate evolution signals rather than allowing abrupt profile jumps; (c) *Directional Consistency*: consecutive profile updates and tracker accumulations are logically coherent in direction; when tracker signals accumulate sufficiently, conversion to profile updates is natural and traceable.

E.4.3 Dim III: Environmental Grounding

Evaluates whether the character truly "lives" in the current scene rather than conversing in a vacuum, reflecting the constraining effect of the World Model on Character Agent behavior.

Environment Awareness (EA). Whether the character's behavior is constrained by the current environment (global world state + location states).

Non-Environment Character Agent: (a) *Global Awareness*: reacts reasonably to the global state (e.g., tension during wartime, conservation during resource scarcity); (b) *Location Awareness*: notices the current state of the location (e.g., a shopkeeper mentioning "last night's storm blew the roof off"); (c) *State Change Response*: when world state changes between scenes, the character notices and adjusts accordingly.

Environment Character Agent: (a) *Global State Consistency*: environmental descriptions are consistent with the current global state; (b) *Location State Accuracy*: environmental descriptions accurately reflect the location card's current state (e.g., damaged buildings not described as intact); (c) *State Change Presentation*: when world state changes between scenes, environmental descriptions reflect these transitions.

Environmental Utilization (EU). Whether characters appropriately and meaningfully use environmental elements when relevant to the scene.

Non-Environment Character Agent: (a) *Environmental Sensory Details*: sensory descriptions convey the character's perception of their environment rather than generic descriptions; (b) *Prop Interaction*: items and entities in the location are used to advance the plot; (c) *Atmosphere Building*: environmental atmosphere enhances immersion rather than conversing in a "blank room."

Environment Character Agent: (a) *Multi-Sensory Richness*: environmental descriptions engage multiple sensory dimensions (visual, auditory, etc.); (b) *Scene Element Usage*: descriptions utilize items and entities in the location; (c) *Atmosphere-Narrative Alignment*: atmosphere matches the current narrative pace and emotional tone.

E.4.4 Dim IV: Interaction Quality

Evaluates whether the character truly "listens" and "responds" to other characters, and whether interactions drive narrative development.

Contextual Responsiveness (CR). Whether responses closely follow context, and whether inter-character attitudes match relationship settings and adjust as profiles evolve. Criteria: (a) *Information Continuity*: not ignoring key information or questions, not abruptly changing topics; (b) *Logical Continuity*: reacting reasonably to others' actions (e.g., accepting/refusing a handed item rather than ignoring it); (c) *Relationship Matching*: clear distinctions in tone and trust toward allies, enemies, and strangers, dynamically adjusting with the plot.

Narrative Progression (NP). Whether interactions advance the narrative and whether previously planted foreshadowing is followed up on. Criteria: (a) *Information Increment*: each round provides new information, actions, or emotional developments rather than repeating known content; (b) *Suspense and Hooks*: suspense created through silence, conflict, or hints, leaving hooks for subsequent interactions; (c) *Foreshadowing Payoff*: foreshadowing from previous scenes is noticed and followed up at appropriate moments.

E.4.5 Dim V: Motivation Generation

Motivation Quality (MQ). Whether the generated scene motivation matches the current profile, world state, and scene settings, and is specific and actionable. Criteria: (a) *Profile Alignment*: motivation aligns with the character's current personality, goals, and relationship status; (b) *Situational Fit*: motivation considers the current world state and scene settings (e.g., no leisure motivations in dangerous scenarios); (c) *Actionability*: motivation is specific enough to guide behavior in the scene

CC: Character Consistency **EQ**: Evolution Quality **EG**: Environmental Grounding **IQ**: Interaction Quality **MG**: Motivation Generation
IC: Instruction Compliance **PF**: Profile Fidelity **SSF**: Speaking Style Fidelity **MDB**: Motivation-Driven Behavior **PUF**: Profile Update Fidelity
PES: Profile Evolution Smoothness **EA**: Environment Awareness **EU**: Environmental Utilization
CR: Contextual Responsiveness **NP**: Narrative Progression **MQ**: Motivation Quality

Models	CC			EQ		EG		IQ		MG	IC	Avg.
	PF	SSF	MDB	PUF	PES	EA	EU	CR	NP	MQ	IC	
Closed-source												
Kimi-K2.5	80.79±10.84	61.45±19.00	87.52 ±9.78	81.07 ±8.63	54.51±16.38	81.29±10.59	83.86±13.00	88.31 ±8.27	56.02±16.66	77.12±12.25	82.83 ±5.68	75.89
Gemini-2.5-Flash	75.34 ±3.94	61.68 ±6.80	75.67 ±5.96	68.61 ±5.91	58.02 ±7.68	67.02 ±6.14	57.10 ±8.90	75.72 ±4.02	51.86 ±9.17	74.61 ±8.04	51.45±26.93	65.19
Gemini-2.5-Pro	93.88 ±3.47	85.94 ±5.90	94.95 ±3.44	83.28 ±5.33	70.51 ±8.65	89.53 ±5.19	90.12 ±6.15	93.63 ±3.25	72.10 ±8.19	77.20 ±6.36	86.06 ±3.08	85.20
Gemini-3.1-Pro-P	89.79 ±5.23	79.40 ±6.91	93.09 ±3.57	78.90 ±5.87	66.29 ±9.98	85.35 ±6.17	83.90 ±8.33	91.87 ±3.62	67.74 ±9.63	83.04 ±6.41	84.12±12.90	82.14
GPT-4o	71.45 ±6.84	54.13 ±9.15	76.36 ±6.76	57.66 ±8.60	54.46 ±8.62	75.80 ±7.04	76.32 ±9.16	77.94 ±4.63	55.89±11.73	69.94 ±8.81	74.15±16.39	67.65
GPT-5-Chat	80.73 ±5.60	73.44 ±9.14	85.89 ±5.45	71.33 ±7.30	64.00 ±9.68	86.74 ±5.29	92.52 ±6.49	87.36 ±4.09	67.37 ±9.19	81.49 ±7.85	77.29±20.63	78.92
GPT-5.1-Chat	81.04 ±9.50	69.93±11.17	75.27 ±8.81	71.55 ±4.89	70.94±10.26	73.41±10.56	71.94±10.62	77.37 ±8.01	37.08±10.01	76.03 ±9.84	81.37 ±5.57	71.45
GPT-5.3-Chat	94.71 ±2.75	88.77 ±4.30	93.71 ±2.95	77.90 ±3.74	80.16 ±6.62	91.62 ±4.88	92.10 ±7.54	92.27 ±3.74	59.98 ±7.55	83.94 ±6.50	83.85 ±3.25	85.36
Claude-4.6-Sonnet	96.47 ±4.16	95.52 ±4.51	98.49 ±3.29	95.30 ±5.61	71.02±15.19	94.33 ±5.25	96.79 ±7.27	97.29 ±4.67	84.03 ±9.31	90.17 ±5.07	62.07±35.76	89.23
Claude-4.6-Opus	**97.81** ±1.92	**96.70** ±2.79	**99.00** ±1.63	**96.36** ±2.83	**84.62** ±9.15	**95.91** ±3.36	**98.73** ±2.46	**98.85** ±1.49	**89.76** ±6.06	**95.36** ±3.63	**91.53** ±3.25	**94.97**
Open-source												
$<$ **14B**												
Qwen3-4B-I	15.31±10.34	7.25 ±6.89	13.57±10.80	21.84±10.54	26.89±11.14	30.11±11.88	32.68±10.21	15.69±12.49	12.47 ±9.73	39.00±18.49	24.76±17.68	21.77
Qwen-4B (EW-B)	43.30±11.99	36.44±11.02	35.73±10.17	35.66 ±8.87	30.61±10.23	41.26 ±8.01	28.30 ±5.43	36.58 ±9.19	26.90 ±7.36	40.02 ±9.04	56.04 ±9.21	37.35
Qwen-4B (EW-F)	44.68±10.21	37.24 ±9.73	37.10 ±8.45	35.24 ±7.81	31.43±10.20	40.09 ±6.84	28.15 ±5.41	35.64 ±9.28	28.30 ±5.88	40.57 ±9.72	59.12±11.15	37.96
Qwen2.5-7B-I	15.77 ±6.68	7.69 ±5.06	11.70 ±6.98	32.25 ±7.42	25.07±10.49	30.45 ±6.41	28.53 ±6.75	11.80 ±7.31	9.69 ±5.53	36.20±16.06	20.14±12.72	20.84
Qwen-7B (Coser)	16.59±11.97	12.20 ±8.33	27.61±12.64	18.16±16.58	17.91±16.25	37.34 ±8.69	25.20 ±4.14	13.92 ±6.96	27.60±12.86	2.06 ±9.94	10.26 ±1.55	18.98
Qwen-7B (Crab)	14.25 ±7.83	6.34 ±4.43	15.76 ±7.80	19.82 ±7.22	15.89 ±9.94	32.79±10.41	20.28 ±5.85	16.91 ±8.45	17.74 ±5.31	26.82±12.05	14.91 ±8.10	18.32
Qwen-7B (EW-B)	50.87 ±9.33	45.47 ±8.89	43.31 ±8.78	40.13 ±6.47	37.89±11.31	43.00 ±7.12	29.83 ±5.03	46.09 ±8.21	32.98 ±6.45	42.05 ±7.76	65.16 ±9.52	43.34
Qwen-7B (EW-F)	**54.07** ±9.85	**48.19**±10.44	**47.15** ±8.44	41.24 ±7.72	40.33±10.06	45.44 ±7.27	31.81 ±6.14	**48.61** ±9.92	36.09 ±6.68	42.12 ±8.31	**65.78** ±9.45	45.53
Llama-3.1-8B-I	30.58±11.10	22.30 ±8.36	30.71±12.65	21.87±16.04	20.36±13.89	44.33 ±8.88	**46.19** ±9.30	40.78±12.36	**40.77** ±9.04	27.26±22.46	14.55±17.16	30.88
Llama-8B (Coser)	30.36±16.76	22.26±15.53	38.28±18.63	15.86±16.21	15.94±15.91	37.61 ±8.57	31.04 ±2.85	30.36±14.55	37.27±11.12	4.07±14.52	11.09 ±4.60	24.92
Llama-8B (Crab)	29.37±18.52	20.00±16.45	38.10±20.05	1.26 ±5.04	2.27 ±8.62	33.20 ±5.08	32.67 ±2.56	26.57±12.04	34.03 ±7.19	10.45±20.33	10.25 ±1.88	21.65
Llama-8B (EW-B)	51.23±19.86	44.21±19.17	45.01±17.42	42.15±10.44	42.09±10.77	45.01±12.20	32.73 ±7.00	46.21±20.25	37.88±14.49	47.31 ±8.00	61.94±15.74	45.07
Llama-8B (EW-F)	50.94±20.32	44.90±18.45	46.43±16.41	**42.49**±10.11	**42.44**±10.66	**46.47**±11.92	33.75 ±7.68	46.03±19.69	38.83±13.13	**47.96** ±7.72	65.62 ±8.79	**45.99**
$\geq$ **14B**												
Qwen2.5-14B-I	31.44±10.64	14.71 ±7.63	30.43±10.11	43.95 ±6.51	35.56±11.66	43.84 ±8.65	38.79 ±9.18	32.83 ±9.13	15.66 ±6.22	53.36±11.16	39.78±18.50	34.58
Qwen-14B (EW-B)	63.50±13.26	56.06±12.60	55.73±11.52	53.17 ±8.10	56.68 ±9.98	50.68 ±8.45	35.47 ±6.55	57.60±13.80	44.90 ±8.74	48.39 ±7.69	**69.32**±10.20	53.77
Qwen-14B (EW-F)	62.39±14.45	55.74±13.59	55.63±13.02	49.82 ±8.48	53.60±11.17	50.03 ±9.79	36.25 ±7.47	55.36±14.26	44.04 ±9.66	49.85 ±7.98	66.27±14.59	52.63
Qwen2.5-32B-I	19.49±11.17	7.42 ±6.16	20.08 ±9.81	35.85 ±9.25	29.82±11.19	39.50 ±8.48	32.63 ±8.40	24.55 ±8.20	13.87 ±6.75	45.12±12.65	38.14±15.37	27.86
Qwen-32B (EW-B)	67.80±11.60	**60.68**±12.84	61.76±10.93	**57.95** ±7.89	**59.52** ±9.81	53.70 ±7.73	37.68 ±6.59	61.25±12.24	47.75 ±9.18	54.51 ±8.48	65.39±19.91	57.09
Qwen-32B (EW-F)	**67.81**±12.50	60.56±12.20	61.34±10.76	57.24 ±7.83	58.62±10.76	55.43 ±9.28	39.22 ±7.91	60.55±13.42	48.48 ±9.52	54.93 ±7.91	63.45±22.36	57.06
Qwen2.5-72B-I	25.71±10.42	9.55 ±5.89	30.88±10.99	37.77 ±8.08	38.90 ±6.64	46.31 ±7.83	36.60 ±7.27	36.50±14.50	23.44±10.29	43.86 ±9.17	41.56±32.31	33.73
Qwen3-32B	42.16±13.59	26.76±12.91	58.39±10.81	51.10 ±9.07	35.24±12.76	62.37±10.82	69.64±12.52	65.99 ±9.24	47.34±11.69	70.18 ±9.74	54.14±19.77	53.03
Llama-3.1-70B-I	32.95±10.28	14.42 ±7.61	39.02±12.63	22.42±17.06	23.05±17.16	46.35 ±9.10	37.78±10.11	48.64±12.57	26.93±10.26	48.42±11.14	38.27±29.68	34.39
Llama-3.3-70B-I	22.66 ±8.93	6.23 ±4.10	27.03±11.61	36.36±10.95	40.42 ±5.32	46.90 ±7.07	39.27 ±8.92	39.83±11.62	22.21±12.14	52.85±11.61	46.33±28.49	34.55
Mistral-Small	54.19 ±9.61	39.24 ±9.77	57.23 ±7.62	46.02 ±5.48	44.53 ±8.14	68.46 ±6.27	70.89 ±7.01	61.67 ±6.76	40.48 ±6.78	52.40 ±9.10	63.85 ±5.01	54.45
DeepSeek-V3-0324	66.22±13.05	56.23±12.75	**75.59** ±8.32	55.73 ±7.73	53.28±11.00	**71.96** ±9.04	**76.43**±10.32	**72.48** ±8.38	**51.89**±14.65	**70.20** ±8.05	55.04±16.32	**64.10**

Table 11: Full benchmark results for **Character Agent** evaluation. I and P denote Instruct and Preview in model names. The best performances within the same model scale are **bold-faced**, and the second-best are underlined.

rather than being vague and abstract.

E.4.6 Dim VI: Instruction Compliance

Instruction Compliance (IC). Whether the character strictly plays itself without overstepping and the output format is standardized. Criteria: (a) *No Overstepping*: strictly outputting only one's own character content, not speaking or acting for others; (b) *Format Compliance*: correct usage of thought/action/speech, output structure meets requirements; (c) *Length Control*: reasonable output length, neither excessively verbose nor overly brief. *Error Penalty.* When a simulation terminates prematurely due to the Character Agent failing to produce valid output, both the IC Penalty and the Metric Penalty described in §E are applied.

E.5 World Model Evaluation (WORLD Score)

E.5.1 Dim I: Scene Planning

Evaluates whether the World Model's scene planning (Liu et al., 2026a,c) is reasonable, including character selection, location/scenario generation, and cross-scene coherence.

Cast Selection Rationality (CSR). Whether the selected participating characters match the current narrative state and character goals. Criteria: (a) *Narrative-Driven*: appearing characters serve the current narrative needs (e.g., opposing characters appear together in conflict scenes); (b) *Goal Relevance*: selected characters are directly related to the current narrative thread; (c) *Avoid Redundancy*: no characters unrelated to the current narrative are introduced; (d) *No Missing Key Characters*: characters in the pool who should appear are not overlooked.

Location & Scenario Rationality (LSR). Whether the chosen location and generated scenario are appropriate for the selected cast and current narrative state. Criteria: (a) *Location Appropriateness*: the location is a plausible place for the selected characters to meet, serving narrative needs; (b) *Scenario Quality*: the scenario provides a clear dramatic setup that is specific and actionable; (c) *Continuity with Previous Scene*: location/scenario follows naturally from the previous scene's events;

SP: Scene Planning **SM**: Speaker Management **WSM**: World State Maintenance **IC**: Instruction Compliance
CSR: Cast Selection Rationality **LSR**: Location & Scenario Rationality **SCC**: Scene Continuity & Coherence
TSO: Turn & Scene Orchestration **GUS**: Global Update Sensitivity **GSA**: Global State Accuracy
LUS: Location Update Sensitivity **LSA**: Location State Accuracy

Models	SP			SM	WSM				IC	Avg.
	CSR	LSR	SCC	TSO	GUS	GSA	LUS	LSA	IC	
Closed-source										
Kimi-K2.5	78.36 ±7.78	80.61±15.01	51.72±25.95	58.84±10.95	66.64 ±7.53	**59.85** ±8.22	68.90 ±6.87	70.54 ±9.77	77.94 ±9.94	68.16
Gemini-2.5-Flash	73.18 ±6.95	84.09 ±6.10	66.36±14.98	43.02±24.56	60.41 ±7.49	56.78 ±4.10	54.22±11.27	56.25 ±6.31	43.49±23.83	59.76
Gemini-2.5-Pro	78.86 ±4.73	88.33 ±6.27	70.34±20.14	75.66 ±5.50	65.89 ±4.69	56.99 ±5.84	54.96 ±8.66	68.20 ±9.76	80.36 ±7.15	71.07
Gemini-3.1-Pro-P	83.50 ±3.17	91.60 ±4.81	69.16±20.97	72.70±10.18	65.60 ±9.89	55.13±10.68	66.51 ±9.67	61.54±10.84	85.55±13.49	72.37
GPT-4o	77.80 ±8.81	85.33 ±6.55	51.02±16.90	61.44 ±7.34	61.40 ±8.32	52.86 ±5.91	54.12 ±9.96	51.44 ±7.07	73.50±18.44	63.21
GPT-5-Chat	80.44 ±6.88	88.30 ±6.12	68.38±16.61	64.69±17.27	60.57 ±9.14	53.97 ±6.21	52.59 ±7.83	57.85 ±7.87	77.51±22.07	67.14
GPT-5.1-Chat	79.61 ±5.63	85.17 ±7.68	42.52±19.01	50.54 ±9.47	68.69 ±2.01	53.83 ±4.19	69.82 ±3.48	53.06 ±6.28	81.69 ±4.83	64.99
GPT-5.3-Chat	80.04 ±5.09	89.35 ±4.81	68.40±15.48	69.04 ±6.41	**70.02** ±2.12	51.11 ±5.48	**75.21** ±6.30	64.77±10.96	85.57 ±4.92	72.61
Claude-4.6-Sonnet	81.89 ±5.46	92.90 ±4.58	60.77±25.33	79.84±13.54	38.93±24.04	33.26±20.37	30.33±21.21	41.12±26.94	58.83±34.31	57.54
Claude-4.6-Opus	**84.86** ±5.34	**96.66** ±2.56	**82.28**±13.37	**84.98** ±4.76	66.05 ±5.74	58.48 ±6.75	61.72 ±9.46	**75.49** ±8.93	**89.34** ±6.50	**77.76**
Open-source										
< 14B										
Qwen3-4B-I	60.20±15.88	52.50±23.75	20.71±15.06	5.65 ±6.64	41.98±23.54	35.56±20.73	22.68±14.07	23.31±13.57	25.78±18.28	32.06
Qwen-4B (EW-B)	61.06±10.75	60.69±12.10	22.63±17.51	29.73 ±6.46	55.63 ±7.21	44.11 ±7.51	57.07 ±6.05	46.06 ±4.64	71.62±11.81	49.84
Qwen-4B (EW-F)	61.78±10.20	62.42±11.14	23.20±16.20	30.81 ±6.65	57.62 ±7.97	44.15 ±8.01	58.09 ±7.59	46.55 ±6.40	72.24±13.19	50.76
Qwen2.5-7B-I	61.07±12.78	51.25±14.71	14.09±14.59	5.82 ±4.92	53.53±11.52	44.53 ±9.69	28.65 ±9.43	31.53 ±8.02	35.58±16.83	36.23
Qwen-7B (EW-B)	61.26±10.29	61.48 ±9.78	29.34±17.50	37.26 ±4.80	**62.62** ±3.69	47.21 ±5.58	**61.38** ±4.10	48.15 ±4.54	75.67 ±9.70	53.82
Qwen-7B (EW-F)	**63.90** ±8.89	63.32±11.66	32.02±17.41	**39.98** ±6.79	61.21 ±5.41	**47.52** ±6.02	61.20 ±4.10	**49.10** ±4.44	**76.59** ±9.78	**54.98**
Llama-3.1-8B-I	60.56±17.34	**68.01**±17.31	34.16±13.44	28.65±11.65	49.02±19.99	43.94±17.78	45.89±19.33	41.12±16.92	30.82±36.53	44.69
Llama-8B (EW-B)	63.78 ±7.75	65.26±12.28	36.07±23.32	38.60±10.93	59.52 ±5.01	44.23 ±6.82	59.53 ±5.59	46.34 ±5.06	71.45±13.73	53.86
Llama-8B (EW-F)	63.58 ±7.91	67.57±11.51	**38.41**±23.29	39.79±11.31	56.59 ±7.62	42.89 ±7.78	60.19 ±5.80	46.75 ±4.99	72.90 ±7.92	54.30
≥ 14B										
Qwen2.5-14B-I	70.66±10.08	67.37±13.98	21.47±15.64	16.95 ±6.74	50.64±21.59	41.28±17.15	23.71±12.53	29.82±13.41	41.89±18.78	40.42
Qwen-14B (EW-B)	68.55 ±8.31	73.50 ±8.80	46.14±16.49	46.56±10.09	61.12 ±5.35	48.62 ±5.52	**62.13** ±4.10	49.73 ±4.68	**79.89**±10.29	59.58
Qwen-14B (EW-F)	68.18 ±7.96	70.67±10.66	39.24±18.05	45.19±10.48	59.45 ±7.84	46.76 ±6.86	60.84 ±6.83	49.08 ±6.37	77.02±15.15	57.38
Qwen2.5-32B-I	69.55 ±9.80	60.18±14.46	17.69±14.67	18.03 ±7.14	61.21 ±6.80	49.53 ±7.00	42.35±13.03	42.08 ±7.58	51.12±17.99	45.75
Qwen-32B (EW-B)	69.58 ±8.40	78.03 ±8.04	43.52±18.12	47.21 ±8.56	61.60 ±3.31	49.68 ±5.28	60.62 ±4.36	50.89 ±4.36	75.47±22.25	59.62
Qwen-32B (EW-F)	69.77 ±7.99	77.26 ±9.58	45.42±17.69	**48.21** ±9.83	61.67 ±3.38	**49.92** ±5.06	60.65 ±4.40	**51.95** ±3.96	73.94±24.58	**59.87**
Qwen2.5-72B-I	70.29 ±9.27	69.52 ±8.44	37.89±13.99	29.88±11.59	39.85±26.68	31.12±20.86	27.04±19.54	29.36±20.08	37.35±29.99	41.37
Qwen3-32B	67.59 ±9.82	71.76±12.88	34.70±20.55	43.60 ±9.59	54.41 ±8.87	47.33 ±7.26	53.52 ±9.16	49.02 ±9.48	51.34±20.40	52.59
Llama-3.1-70B-I	63.89±13.92	72.24±12.37	30.85±15.45	28.53±10.52	50.82±18.89	40.32±15.07	21.66 ±9.96	35.75±13.93	31.37±25.21	41.71
Llama-3.3-70B-I	75.62±10.65	78.39±12.61	36.76±14.19	23.47 ±9.18	41.47±22.18	34.41±18.74	22.41±15.41	27.44±15.72	34.90±22.87	41.65
Mistral-Small	72.24 ±6.72	73.74 ±6.86	41.38±16.00	42.39 ±6.98	60.88 ±7.47	44.55 ±6.82	41.64 ±8.90	44.92 ±6.53	60.41 ±8.50	53.57
DeepSeek-V3-0324	**76.98** ±6.36	**81.86** ±8.43	**51.00**±19.30	38.75±15.05	**63.48** ±5.11	48.79 ±6.70	47.32±10.08	49.70 ±8.14	60.33±16.47	57.58

Table 12: Full benchmark results for **World Model** evaluation. I and P denote Instruct and Preview in model names. The best performances within the same model scale are **bold-faced**, and the second-best are underlined.

(d) *Character-Setting Fit*: the setting is appropriate for the selected characters.

Scene Continuity & Coherence (SCC). Whether cross-scene planning forms a coherent narrative arc (cross-scene metric evaluated over the full trajectory). Criteria: (a) *Narrative Arc*: consecutive scenes form a directional narrative progression rather than random assembly; (b) *Scene Transitions*: location choices and scene descriptions naturally connect with the previous scene; (c) *Pacing Control*: overall narrative pacing is reasonable, important plot points receive sufficient development; (d) *Thread Management*: narrative threads are introduced, developed, and resolved; no plot lines are abandoned.

E.5.2 Dim II: Speaker Management

Turn & Scene Orchestration (TSO). Whether speaker selection, environmental description timing, and scene ending timing are appropriate. Criteria: (a) *Speaker Selection*: `next_character` selects the character who should most appropriately respond, not random rotation; (b) *Environmental Description Timing*: environmental descriptions introduced at appropriate moments (scene changes, important events); (c) *Group Character Actions*: in multi-character scenes, character combinations are reasonably selected for joint interactions, fitting the current situation and relationships; (d) *Character Coverage Balance*: core characters receive participation proportional to their narrative importance; transient characters naturally fade out after fulfilling their narrative function; (e) *Ending Timing*: the scene ends at a natural narrative juncture, not abruptly during a climax or dragging when nothing happens.

E.5.3 Dim III: World State Maintenance*

Evaluates whether the World Model's maintenance of global state and location state is accurate and timely. Unique to this framework, evaluating global state and location state separately.

Global Update Sensitivity (GUS). Whether the timing of global state updates is appropriate. Criteria: (a) *No Over-Updating*: casual conversations or local events should not trigger global state updates; (b) *No Missing Updates*: truly globally impactful events (war, kingdom falling) must be captured; (c) *Trigger Judgment*: correctly distinguishing "local impact" from "global impact" events.

CC: Character Consistency **EQ**: Evolution Quality **EG**: Environmental Grounding **IQ**: Interaction Quality **MG**: Motivation Generation
PF: Profile Fidelity **SSF**: Speaking Style Fidelity **MDB**: Motivation-Driven Behavior **PUF**: Profile Update Fidelity **PES**: Profile Evolution Smoothness
EA: Environment Awareness **EU**: Environmental Utilization **CR**: Contextual Responsiveness **NP**: Narrative Progression **MQ**: Motivation Quality

Models	Setting	CC			EQ		EG		IQ		MG	Avg.
		PF	SSF	MDB	PUF	PES	EA	EU	CR	NP	MQ	
GPT-5.3-Chat	Full	94.71 ±2.75	88.77 ±4.30	93.71 ±2.95	77.90 ±3.74	80.16 ±6.62	91.62 ±4.88	92.10 ±7.54	92.27 ±3.74	59.98 ±7.55	83.94 ±6.50	85.52
	w/o Char State	88.48 ±5.86	86.08 ±5.92	86.54 ±6.60	27.18 ±8.88	25.54±10.35	90.92 ±4.75	91.94 ±7.38	89.89 ±4.52	53.60 ±8.15	57.08±12.97	69.73
	w/o Both	86.38 ±5.77	82.57 ±7.18	83.63 ±6.86	24.96±10.45	25.36 ±9.88	86.39 ±7.18	84.29 ±9.84	86.14 ±5.61	52.61 ±7.36	55.46±12.57	66.78
Llama-3.1-8B-I	Full	30.58±11.10	22.30 ±8.36	30.71±12.65	21.87±16.04	20.36±13.89	44.33 ±8.88	46.19 ±9.30	40.78±12.36	40.77 ±9.04	27.26±22.46	32.52
	w/o Char State	12.31±13.52	11.23±10.97	15.37 ±8.40	6.38±12.87	13.61 ±8.99	32.64±11.83	39.45 ±6.17	19.22±11.23	21.09 ±7.72	23.21±14.15	19.45
	w/o Both	5.78 ±5.94	4.83 ±5.03	7.29 ±5.75	5.68 ±3.58	7.01±12.21	23.08 ±7.68	30.55 ±7.35	11.93 ±7.26	12.12 ±6.44	20.64±24.23	12.89

Table 13: Ablation study on **Character Agent** performance. "Full" denotes the complete EvolvingWorld framework; "w/o Char State" removes character state updates; "w/o Both" removes both world and character state updates.

SP: Scene Planning **SM**: Speaker Management **WSM**: World State Maintenance **CSR**: Cast Selection Rationality
LSR: Location & Scenario Rationality **SCC**: Scene Continuity & Coherence **TSO**: Turn & Scene Orchestration
GUS: Global Update Sensitivity **GSA**: Global State Accuracy **LUS**: Location Update Sensitivity **LSA**: Location State Accuracy

Models	Setting	SP			SM	WSM				Avg.
		CSR	LSR	SCC	TSO	GUS	GSA	LUS	LSA	
GPT-5.3-Chat	Full	80.04 ±5.09	89.35 ±4.81	68.40±15.48	69.04 ±6.41	70.02 ±2.12	51.11 ±5.48	75.21 ±6.30	64.77±10.96	70.99
	w/o World State	78.31 ±4.42	87.63 ±6.28	58.88±16.92	58.65 ±7.10	43.48±16.93	51.87 ±6.00	64.70 ±7.35	44.47 ±8.62	61.00
	w/o Both	79.65 ±4.93	87.81 ±8.68	52.94±17.52	56.18 ±6.84	42.20±20.13	51.48 ±7.21	64.60 ±7.01	43.10 ±8.32	59.75
Llama-3.1-8B-I	Full	60.56±17.34	68.01±17.31	34.16±13.44	28.65±11.65	49.02±19.99	43.94±17.78	45.89±19.33	41.12±16.92	46.42
	w/o World State	51.06±16.65	59.21±14.29	27.02±12.89	5.79 ±5.81	18.82±15.83	20.04±15.71	29.80±18.99	15.27±20.79	28.38
	w/o Both	41.94±11.16	43.51±12.99	7.73±11.71	4.10 ±2.84	20.31±15.33	20.72±11.15	28.45±20.52	17.57±14.14	23.04

Table 14: Ablation study on **World Model** performance. "Full" denotes the complete EvolvingWorld framework; "w/o World State" removes world state updates; "w/o Both" removes both world and character state updates.

Setting	PUF	PES	Avg.
Full	77.90±3.74	80.16±6.62	79.03
w/o Hidden Tracker	65.30±5.92	77.56±4.91	71.43
w/o Open Schema	76.65±2.94	78.91±4.42	77.78

Table 15: Character-side component ablations with *GPT-5.3-Chat*. Removing the Hidden Tracker or replacing the open profile with a fixed schema both degrade profile-update quality.

Setting	CSR	LSR	GUS	GSA	Avg.
Full	80.04±5.09	89.35±4.81	70.02±2.12	51.11±5.48	72.63
w/o Open Schema	79.10±3.44	88.55±2.53	66.50±1.89	50.95±2.15	71.28

Table 16: World-side schema ablation with *GPT-5.3-Chat*. Replacing the open global state with a fixed schema reduces global-state maintenance.

Global State Accuracy (GSA). Whether the updated global state content is accurate. Criteria: (a) *Factual Accuracy*: global state accurately reflects occurred events without erroneous information; (b) *Timely Retirement*: overturned or outdated information is removed or updated; (c) *Concise Expression*: descriptions remain concise without accumulating redundant details.

Location Update Sensitivity (LUS). Whether the timing of location state updates is appropriate. Criteria: (a) *No Over-Updating*: temporary events without lasting impact should not trigger updates; (b) *No Missing Updates*: events with lasting physical or environmental changes must be captured; (c) *Persistence Judgment*: correctly distinguishing "temporary changes" from "persistent changes."

Location State Accuracy (LSA). Whether the updated location state and Important Entities list are accurate. Criteria: (a) *Spatial Consistency*: spatial logic of location descriptions is self-consistent; (b) *Entity Accuracy*: Important Entities list accurately reflects entities currently present; (c) *Cross-Scene Continuity*: descriptions of the same location across scenes remain consistent.

Evaluation Target	Model	Untrained	Trained	
		Avg.	ID Avg.	OOD Avg.
Character Agent	Qwen3-4B-I	21.77	38.80	36.95
	Qwen2.5-7B-I	20.84	45.28	45.88
	Llama-3.1-8B-I	30.88	49.62	41.08
	Qwen2.5-14B-I	34.58	52.57	52.73
	Qwen2.5-32B-I	27.86	56.63	57.61
World Model	Qwen3-4B-I	32.06	52.36	48.84
	Qwen2.5-7B-I	36.23	55.44	54.23
	Llama-3.1-8B-I	44.69	56.73	50.98
	Qwen2.5-14B-I	40.42	58.44	55.91
	Qwen2.5-32B-I	45.75	60.65	58.96

Table 17: ID and OOD performance after full-mixture training on Character Agent and World Model. Untrained averages are included as references.

E.5.4 Dim IV: Instruction Compliance

Instruction Compliance (IC). Whether the output format is correct and the World Model acts within its scope of responsibility. Criteria: (a) *Format Correctness*: JSON format and fields for each task output are complete and correct; (b) *No Overstepping*: the World Model strictly acts within its own responsibility, not generating character dialogue; (c) *Field Completeness*: all required fields are filled without omissions.

Error Penalty. When a simulation terminates prematurely due to the World Model failing to produce valid output, both the IC Penalty and the Metric

CC: Character Consistency **EQ**: Evolution Quality **EG**: Environmental Grounding **IQ**: Interaction Quality **MG**: Motivation Generation
IC: Instruction Compliance **PF**: Profile Fidelity **SSF**: Speaking Style Fidelity **MDB**: Motivation-Driven Behavior **PUF**: Profile Update Fidelity
PES: Profile Evolution Smoothness **EA**: Environment Awareness **EU**: Environmental Utilization
CR: Contextual Responsiveness **NP**: Narrative Progression **MQ**: Motivation Quality

Models	CC			EQ		EG		IQ		MG	IC	Avg.
	PF	SSF	MDB	PUF	PES	EA	EU	CR	NP	MQ	IC	
Claude-4.6-Sonnet Judge												
Claude-4.6-Opus	97.81 ±1.92	96.70 ±2.79	99.00 ±1.63	96.36 ±2.83	84.62 ±9.15	95.91 ±3.36	98.73 ±2.46	98.85 ±1.49	89.76 ±6.06	95.36 ±3.63	91.53 ±3.25	94.97
GPT-5.3-Chat	94.71 ±2.75	88.77 ±4.30	93.71 ±2.95	77.90 ±3.74	80.16 ±6.62	91.62 ±4.88	92.10 ±7.54	92.27 ±3.74	59.98 ±7.55	83.94 ±6.50	83.85 ±3.25	85.36
Gemini-3.1-Pro-P	89.79 ±5.23	79.40 ±6.91	93.09 ±3.57	78.90 ±5.87	66.29 ±9.98	85.35 ±6.17	83.90 ±8.33	91.87 ±3.62	67.74 ±9.63	83.04 ±6.41	84.12±12.90	82.14
Kimi-K2.5	80.79±10.84	61.45±19.00	87.52 ±9.78	81.07 ±8.63	54.51±16.38	81.29±10.59	83.86±13.00	88.31 ±8.27	56.02±16.66	77.12±12.25	82.83 ±5.68	75.89
GPT-4o	71.45 ±6.84	54.13 ±9.15	76.36 ±6.76	57.66 ±8.60	54.46 ±8.62	75.80 ±7.04	76.32 ±9.16	77.94 ±4.63	55.89±11.73	69.94 ±8.81	74.15±16.39	67.65
DeepSeek-V3-0324	66.22±13.05	56.23±12.75	75.59 ±8.32	55.73 ±7.73	53.28±11.00	71.96 ±9.04	76.43±10.32	72.48 ±8.38	51.89±14.65	70.20 ±8.05	55.04±16.32	64.10
Qwen-32B (EW-F)	67.81±12.50	60.56±12.20	61.34±10.76	57.24 ±7.83	58.62±10.76	55.43 ±9.28	39.22 ±7.91	60.55±13.42	48.48 ±9.52	54.93 ±7.91	63.45±22.36	57.06
Mistral-Small	54.19 ±9.61	39.24 ±9.77	57.23 ±7.62	46.02 ±5.48	44.53 ±8.14	68.46 ±6.27	70.89 ±7.01	61.67 ±6.76	40.48 ±6.78	52.40 ±9.10	63.85 ±5.01	54.45
Qwen-14B (EW-F)	62.39±14.45	55.74±13.59	55.63±13.02	49.82 ±8.48	53.60±11.17	50.03 ±9.79	36.25 ±7.47	55.36±14.26	44.04 ±9.66	53.36±11.16	66.27±14.59	52.95
Qwen-7B (EW-F)	54.07 ±9.85	48.19±10.44	47.15 ±8.44	41.24 ±7.72	40.33±10.06	45.44 ±7.27	31.81 ±6.14	48.61 ±9.92	36.09 ±6.68	42.12 ±8.31	65.78 ±9.45	45.53
Qwen2.5-14B-I	31.44±10.64	14.71 ±7.63	30.43±10.11	43.95 ±6.51	35.56±11.66	43.84 ±8.65	38.79 ±9.18	32.83 ±9.13	15.66 ±6.22	49.85 ±7.98	39.78±18.50	34.26
Qwen2.5-32B-I	19.49±11.17	7.42 ±6.16	20.08 ±9.81	35.85 ±9.25	29.82±11.19	39.50 ±8.48	32.63 ±8.40	24.55 ±8.20	13.87 ±6.75	45.12±12.65	38.14±15.37	27.86
Qwen2.5-7B-I	15.77 ±6.68	7.69 ±5.06	11.70 ±6.98	32.25 ±7.42	25.07±10.49	30.45 ±6.41	28.53 ±6.75	11.80 ±7.31	9.69 ±5.53	36.20±16.06	20.14±12.72	20.84
Gemini-2.5-Pro Judge												
Claude-4.6-Opus	84.65 ±4.23	80.45 ±7.44	92.07 ±3.36	86.92 ±3.98	79.68 ±7.17	88.46 ±4.20	89.38 ±3.99	91.09 ±3.40	86.53 ±3.20	84.50 ±3.24	64.04 ±4.29	84.34
GPT-5.3-Chat	85.73 ±3.30	82.87 ±5.36	89.30 ±2.81	75.39 ±4.67	77.30 ±6.04	88.39 ±3.83	88.41 ±5.87	88.31 ±2.93	73.48 ±6.61	79.71 ±4.31	72.01 ±3.21	81.90
Gemini-3.1-Pro-P	83.60 ±4.92	79.13 ±7.81	88.93 ±3.25	79.26 ±4.93	73.80 ±8.97	83.29 ±4.41	83.16 ±5.23	86.66 ±3.59	78.22 ±5.58	81.22 ±4.83	73.42±11.74	80.97
Kimi-K2.5	83.36 ±4.52	65.20±13.79	91.10 ±3.02	82.50 ±5.11	74.97 ±7.04	86.26 ±4.70	87.60 ±3.94	87.90 ±4.73	77.37 ±7.42	82.68 ±3.78	63.69 ±6.71	80.24
GPT-4o	79.47 ±4.49	72.33±10.33	86.74 ±3.49	64.29 ±9.81	62.89±10.72	79.41 ±5.93	81.52 ±7.35	79.78 ±4.63	68.53 ±9.50	76.36 ±6.16	60.28±13.10	73.78
DeepSeek-V3-0324	76.21 ±6.86	70.51±10.72	84.34 ±5.69	60.61 ±8.99	62.68 ±9.68	79.85 ±5.65	81.84 ±6.76	69.09 ±9.18	64.16±11.04	77.75 ±6.07	48.10±14.71	70.47
Mistral-Small	69.87 ±7.06	61.54 ±9.15	76.86 ±5.72	52.82 ±6.84	54.97 ±8.92	77.37 ±5.41	84.20 ±4.24	60.79 ±7.67	53.46 ±6.47	68.34 ±7.08	49.25 ±6.69	64.50
Qwen-32B (EW-F)	59.80±12.23	59.11±13.08	63.20±12.35	64.66 ±9.27	64.97±11.64	57.35 ±9.26	43.79 ±9.60	53.50±13.01	52.86 ±9.26	71.36 ±6.86	55.22±22.03	58.71
Qwen-14B (EW-F)	55.79±13.23	56.10±12.47	59.67±12.60	58.05 ±9.74	59.88±11.24	61.10 ±8.13	42.41 ±8.82	49.59±12.53	49.92 ±8.67	69.18 ±8.06	59.81±15.58	56.50
Qwen-7B (EW-F)	49.60±10.42	50.63±10.35	53.17 ±9.29	50.98 ±8.79	51.31±10.88	47.81 ±8.10	39.32 ±6.66	42.38 ±9.19	44.06 ±7.40	65.12 ±7.94	60.29±12.58	50.42
Qwen2.5-14B-I	47.49 ±9.73	34.93 ±9.39	55.31 ±9.53	54.36 ±9.25	51.94±11.02	54.19 ±8.92	47.42±11.09	41.00 ±7.11	33.45 ±7.05	68.23 ±9.21	37.60±18.63	47.81
Qwen2.5-32B-I	40.81 ±9.85	26.48 ±7.61	51.41±10.66	50.30 ±8.55	52.99±10.69	57.30 ±8.26	46.57 ±9.69	42.39 ±8.14	32.81 ±6.77	65.87 ±8.80	36.38±14.19	45.76
Qwen2.5-7B-I	35.03 ±7.12	29.50 ±6.67	42.27 ±7.71	39.44 ±8.88	38.88 ±8.58	47.83 ±7.02	41.93 ±9.51	29.21 ±6.54	28.05 ±5.17	52.62±17.56	25.37±13.48	37.28
GPT-5.1-Chat Judge												
Claude-4.6-Opus	93.08 ±4.07	87.00 ±7.84	97.43 ±2.52	80.97±10.70	81.42 ±8.48	93.31 ±4.59	92.54 ±5.72	95.42 ±2.85	89.45 ±5.51	89.50 ±4.73	63.83±10.68	87.63
GPT-5.3-Chat	93.93 ±3.95	89.38 ±4.85	93.20 ±3.51	75.36 ±6.54	81.52 ±4.97	91.65 ±6.45	88.97±11.99	90.46 ±2.87	63.82±10.64	81.21 ±5.32	62.12 ±5.78	82.87
Gemini-3.1-Pro-P	89.33 ±5.52	77.65±12.23	91.58 ±3.67	69.95 ±8.46	71.13 ±7.19	81.23 ±7.92	75.67±11.12	87.92 ±4.12	72.49±10.98	80.96 ±6.67	62.85±12.29	78.25
GPT-4o	82.65 ±6.41	72.47±13.62	86.25 ±5.47	55.34±11.85	66.73 ±8.78	80.17 ±8.34	79.33±11.01	80.68 ±3.21	63.14±13.16	74.43 ±9.67	48.91±12.32	71.83
Kimi-K2.5	81.63 ±9.13	51.81±21.47	90.70 ±4.90	57.81±11.60	54.20±17.82	83.54 ±8.69	84.18 ±8.53	87.42 ±4.66	67.77±14.99	78.64 ±6.52	42.85±12.99	70.96
DeepSeek-V3-0324	74.51±11.46	59.83±19.41	85.28 ±5.36	52.37±12.58	61.47±12.67	77.05±11.41	78.01±12.23	78.82 ±5.01	64.42±14.64	77.11 ±9.17	32.47±13.94	67.39
Mistral-Small	68.11 ±9.92	43.64±14.56	73.56 ±7.11	49.37 ±8.00	63.48 ±7.51	81.58 ±7.16	82.79 ±8.10	71.23 ±5.02	53.21 ±9.14	58.62±10.79	38.54 ±6.82	62.19
Qwen-32B (EW-F)	64.74±13.62	57.45±12.20	64.08±11.40	42.90±11.75	62.89 ±8.00	57.32±12.48	44.93 ±9.53	65.80±10.37	55.89±10.51	64.03 ±8.61	51.10±21.57	57.38
Qwen-14B (EW-F)	59.28±15.69	52.16±13.53	60.09±11.25	51.83±10.55	62.41 ±7.04	59.63±14.29	41.35 ±8.32	61.47±11.91	52.11 ±9.82	59.61 ±9.91	55.20±15.68	55.92
Qwen-7B (EW-F)	53.47±11.21	42.70±10.81	52.83 ±9.70	31.90±10.49	53.58 ±8.71	44.88±11.78	39.14 ±7.08	56.64 ±8.11	45.50 ±7.84	53.92 ±9.69	55.96±12.78	48.23
Qwen2.5-14B-I	45.67±14.10	23.96±11.81	53.42±11.87	30.42±10.72	47.84±13.46	41.44±11.35	45.88±14.28	52.11 ±9.77	23.05 ±7.55	58.44±13.63	30.09±15.99	41.12
Qwen2.5-32B-I	37.63±15.72	16.16±10.79	42.91±15.04	37.02±10.29	54.83±12.93	47.41±12.67	48.91±13.57	50.08±10.71	23.81 ±8.01	56.38±14.54	31.85±14.74	40.64
Qwen2.5-7B-I	32.51±12.05	15.01 ±7.78	34.43±11.35	24.62 ±8.84	32.14±12.72	36.76 ±8.78	41.13±11.94	26.64 ±8.80	17.98 ±6.01	44.23±18.47	18.16±12.33	29.42

Table 18: Full **Character Agent** benchmark results under three judge models. Gray rows indicate the judge model; all score rows are copied from the original per-judge tables, which are retained separately.

Penalty described in §E are applied.

F Full Results on EvolvingWorld Benchmark

Training mixture settings. Our main experiments use the full EW data mixture (EW-F), which preserves the natural task distribution of the constructed training set. For comparison, we also train a balanced-mixture variant (EW-B), where each task contributes the same number of training examples. Tables 11 and 12 provide the complete EvolvingWorld benchmark results on 21 models and their fine-tuned ones.

Analysis. Across both Character Agent and World Model evaluations, EW-trained models consistently improve over their corresponding open-source backbones and role-playing-only baselines. The full and balanced mixtures show similar overall trends, with the better setting varying by backbone and evaluation role; we therefore use the full mixture as the default setting in the main text and report the balanced variant here for completeness.

G Ablation Study

We ablate the two core mechanisms in EvolvingWorld: character state and world state updates, using *GPT-5.3-Chat* and *Llama-3.1-8B-Instruct*.

Tables 13 and 14 show that both mechanisms matter but affect different parts of the simulation. Removing character state updates sharply reduces character evolution metrics such as PUF, PES, and MQ, causing large character-agent drops for both backbones. Removing world state updates mainly hurts world-model metrics, including scene continuity, turn/scene organization, and state-maintenance. The two mechanisms are also coupled: removing world updates weakens character-side grounding and interaction, while removing character updates further harms world-side continuity and orchestration, indicating that long-horizon

SP: Scene Planning **SM**: Speaker Management **WSM**: World State Maintenance **IC**: Instruction Compliance
CSR: Cast Selection Rationality **LSR**: Location & Scenario Rationality **SCC**: Scene Continuity & Coherence **TSO**: Turn & Scene Orchestration
GUS: Global Update Sensitivity **GSA**: Global State Accuracy **LUS**: Location Update Sensitivity **LSA**: Location State Accuracy

Models	SP			SM	WSM				IC	Avg.
	CSR	LSR	SCC	TSO	GUS	GSA	LUS	LSA	IC	
Claude-4.6-Sonnet Judge										
Claude-4.6-Opus	84.86 ±5.34	96.66 ±2.56	82.28±13.37	84.98 ±4.76	66.05 ±5.74	58.48 ±6.75	61.72 ±9.46	75.49 ±8.93	89.34 ±6.50	77.76
GPT-5.3-Chat	80.04 ±5.09	89.35 ±4.81	68.40±15.48	69.04 ±6.41	70.02 ±2.12	51.11 ±5.48	75.21 ±6.30	64.77±10.96	85.57 ±4.92	72.61
Gemini-3.1-Pro-P	83.50 ±3.17	91.60 ±4.81	69.16±20.97	72.70±10.18	65.60 ±9.89	55.13±10.68	66.51 ±9.67	61.54±10.84	85.55±13.49	72.37
Kimi-K2.5	78.36 ±7.78	80.61±15.01	51.72±25.95	58.84±10.95	66.64 ±7.53	59.85 ±8.22	68.90 ±6.87	70.54 ±9.77	77.94 ±9.94	68.16
GPT-4o	77.80 ±8.81	85.33 ±6.55	51.02±16.90	61.44 ±7.34	61.40 ±8.32	52.86 ±5.91	54.12 ±9.96	51.44 ±7.07	73.50±18.44	63.21
Qwen-32B (EW-F)	69.77 ±7.99	77.26 ±9.58	45.42±17.69	48.21 ±9.83	61.67 ±3.38	49.92 ±5.06	60.65 ±4.40	51.95 ±3.96	73.94±24.58	59.87
DeepSeek-V3-0324	76.98 ±6.36	81.86 ±8.43	51.00±19.30	38.75±15.05	63.48 ±5.11	48.79 ±6.70	47.32±10.08	49.70 ±8.14	60.33±16.47	57.58
Qwen-14B (EW-F)	68.18 ±7.96	70.67±10.66	39.24±18.05	45.19±10.48	59.45 ±7.84	46.76 ±6.86	60.84 ±6.83	49.08 ±6.37	77.02±15.15	57.38
Qwen-7B (EW-F)	63.90 ±8.89	63.32±11.66	32.02±17.41	39.98 ±6.79	61.21 ±5.41	47.52 ±6.02	61.20 ±4.10	49.10 ±4.44	76.59 ±9.78	54.98
Mistral-Small	72.24 ±6.72	73.74 ±6.86	41.38±16.00	42.39 ±6.98	60.88 ±7.47	44.55 ±6.82	41.64 ±8.90	44.92 ±6.53	60.41 ±8.50	53.57
Qwen2.5-32B-I	69.55 ±9.80	60.18±14.46	17.69±14.67	18.03 ±7.14	61.21 ±6.80	49.53 ±7.00	42.35±13.03	42.08 ±7.58	51.12±17.99	45.75
Qwen2.5-14B-I	70.66±10.08	67.37±13.98	21.47±15.64	16.95 ±6.74	50.64±21.59	41.28±17.15	23.71±12.53	29.82±13.41	41.89±18.78	40.42
Qwen2.5-7B-I	61.07±12.78	51.25±14.71	14.09±14.59	5.82 ±4.92	53.53±11.52	44.53 ±9.69	28.65 ±9.43	31.53 ±8.02	35.58±16.83	36.23
Gemini-2.5-Pro Judge										
Claude-4.6-Opus	76.83 ±4.21	84.21 ±2.91	76.99±14.25	83.68 ±4.88	54.78 ±8.51	54.04 ±9.69	44.52 ±8.75	76.89 ±7.02	62.69 ±4.18	68.29
GPT-5.3-Chat	76.63 ±2.94	82.79 ±3.70	70.25±18.19	77.58±10.20	57.94 ±9.40	49.54±11.17	58.92 ±9.80	58.87±11.32	59.67 ±9.48	65.80
Gemini-3.1-Pro-P	76.40 ±3.17	81.22 ±5.97	62.39±16.62	78.08 ±5.82	59.69 ±9.27	47.25±11.46	60.65 ±7.69	60.08 ±9.34	59.77 ±4.23	65.06
Kimi-K2.5	74.55 ±3.94	80.72 ±5.05	66.24±19.49	78.22 ±5.14	60.29 ±5.65	43.17 ±6.30	61.23 ±7.98	56.41±12.61	60.88 ±2.91	64.63
GPT-4o	72.17 ±7.06	80.52 ±4.49	59.92±18.48	72.19 ±5.82	57.93 ±6.73	49.92 ±5.83	51.67 ±7.83	48.83 ±8.40	54.05±11.94	60.80
Qwen-32B (EW-F)	71.63 ±6.42	75.19 ±6.69	44.57±15.47	59.32 ±9.97	58.53 ±7.48	47.17 ±6.12	55.43 ±5.35	47.57 ±6.05	51.88±19.64	56.81
Qwen-14B (EW-F)	69.27 ±9.61	73.98 ±8.09	42.15±16.52	58.99 ±9.15	57.85 ±5.49	45.89 ±7.30	54.98 ±4.66	45.71 ±6.21	55.36±13.36	56.02
DeepSeek-V3-0324	72.79 ±6.30	80.12 ±4.49	63.91±18.77	48.67±16.52	54.00 ±8.16	45.09 ±7.62	43.39 ±7.69	47.63 ±8.13	45.53±13.20	55.68
Qwen-7B (EW-F)	65.21 ±7.91	65.51±12.03	39.13±15.85	55.50 ±7.33	57.02 ±4.43	43.25 ±5.22	53.96 ±4.28	44.25 ±4.81	55.61±11.01	53.27
Mistral-Small	68.50 ±7.89	73.45 ±6.42	45.31±19.01	54.63 ±8.28	54.60 ±6.78	40.76 ±5.81	40.76 ±5.35	37.47 ±6.35	44.67 ±7.17	51.13
Qwen2.5-32B-I	68.32 ±8.02	71.08 ±9.11	44.55±18.29	37.55 ±9.41	57.32 ±5.26	45.31 ±6.05	41.75 ±9.93	38.63 ±8.44	39.56±15.65	49.34
Qwen2.5-14B-I	67.69 ±7.98	66.63±10.02	41.13±19.36	34.07 ±8.22	47.30±20.51	36.99±15.34	28.05±12.87	28.09±13.09	32.80±16.30	42.53
Qwen2.5-7B-I	62.90±11.20	61.27±10.09	34.71±17.83	23.34 ±6.63	54.40±10.54	43.94 ±8.32	35.02 ±8.28	31.56 ±7.73	34.45±16.02	42.40
GPT-5.1-Chat Judge										
Claude-4.6-Opus	74.46 ±4.93	84.15 ±5.61	87.84±10.55	77.36 ±5.15	70.37 ±4.03	61.00 ±4.54	66.50 ±7.22	69.27 ±8.43	84.81 ±7.58	75.08
GPT-5.3-Chat	78.47 ±4.04	86.22 ±4.50	79.80±16.34	76.78±10.79	64.76 ±9.27	59.89 ±7.96	58.88±11.37	64.98 ±9.31	80.54±14.74	72.26
Gemini-3.1-Pro-P	78.78 ±6.56	90.78 ±4.65	83.18±11.77	80.95 ±7.22	63.84 ±7.23	58.76 ±6.11	39.65±11.47	69.03±10.06	70.32±11.77	70.59
Kimi-K2.5	70.18 ±9.64	82.49 ±6.34	67.75±13.53	67.59 ±5.85	63.80 ±8.09	56.94 ±6.08	47.34±13.95	54.09 ±8.46	73.56±19.04	64.86
GPT-4o	76.75 ±5.35	79.72 ±9.81	73.24±17.62	65.47 ±9.63	61.46 ±8.31	58.98 ±6.49	50.13±10.80	61.84 ±9.77	53.25±15.11	64.54
Qwen-32B (EW-F)	69.54 ±8.20	75.29 ±9.79	61.15±20.52	60.45 ±7.52	65.48 ±8.62	55.95 ±8.63	59.63 ±7.36	57.30 ±7.31	71.86±27.66	64.07
Qwen-14B (EW-F)	67.09±10.07	74.38 ±9.20	55.05±21.36	57.83 ±8.72	59.94 ±9.53	54.70 ±7.22	60.69 ±7.78	56.04 ±8.08	76.61±19.13	62.48
DeepSeek-V3-0324	71.48 ±7.03	80.63 ±6.44	69.51±16.65	44.03±16.01	64.00 ±7.78	54.70 ±6.46	52.56±15.91	52.82±13.05	61.68±19.99	61.27
Qwen-7B (EW-F)	62.94 ±8.69	66.13±13.59	43.44±17.70	55.53 ±5.45	63.53 ±5.43	55.68 ±6.52	60.07 ±6.71	55.54 ±7.01	77.26±15.58	60.01
Mistral-Small	65.35 ±8.86	72.99 ±8.19	68.58±17.40	50.67 ±9.02	61.36±10.16	54.76 ±7.42	28.60±11.15	42.50±11.45	50.84±15.99	55.07
Qwen2.5-32B-I	66.21 ±8.56	75.08 ±7.86	52.36±19.60	33.10±10.36	59.29±10.00	54.30 ±6.80	34.88±15.99	44.21±13.90	45.73±23.35	51.68
Qwen2.5-14B-I	66.01 ±9.61	70.85±10.83	46.51±18.19	28.42±10.12	53.14±22.62	46.10±19.56	16.58±10.62	29.33±15.50	35.09±20.99	43.56
Qwen2.5-7B-I	58.42±11.48	64.34±12.99	41.27±23.04	13.45 ±6.19	57.16±14.31	50.50±11.22	23.91±10.16	30.31±11.15	35.35±19.17	41.63

Table 19: Full **World Model** benchmark results under three judge models. Gray rows indicate the judge model; all score rows are copied from the original per-judge tables, which are retained separately.

simulation depends on their joint evolution.

Hidden Tracker. To validate the Hidden Tracker specifically, we remove it from the initial character states, simulation prompts, and character-update outputs, while keeping scene-by-scene profile updates unchanged, using *GPT-5.3-Chat*. As shown in Table 15, removing the Hidden Tracker reduces PUF by 12.60 and PES by 2.60, a 7.60-point drop in their average. This confirms that accumulating weak evidence before committing profile changes is important for deciding when profile dimensions should evolve across scenes.

Open vs. Fixed Schema. To isolate the benefit of the open-schema design, we replace the open schemas with fixed ones while keeping all other settings unchanged, using *GPT-5.3-Chat*. All books share the same fixed schema, where each character profile is compressed into `background`, `personality`, `relationships`, `goals`, and `current_state`, and the global world state into `setting`, `social_rules`, `institutions`, `conflicts`, and `current_events`. Tables 15 and 16 show that fixed schemas consistently reduce the focused averages by 1.25 and 1.35 points, indicating that book-specific dimensions provide a consistent advantage over a single fixed schema.

H In- and Out-of-Distribution Results

We further compare the performance of EvolvingWorld on in-distribution (ID) and out-of-distribution (OOD) test examples. Table 17 reports average scores for both Character Agent and World

Type	Judge Model	1	2	3	4	5	6
Character Agent	Claude-4.6-Sonnet	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	Kimi-K2.5	GPT-4o	DeepSeek-V3-0324
	Gemini-2.5-Pro	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	Kimi-K2.5	GPT-4o	DeepSeek-V3-0324
	GPT-5.1-Chat	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	GPT-4o	Kimi-K2.5	DeepSeek-V3-0324
World Model	Claude-4.6-Sonnet	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	Kimi-K2.5	GPT-4o	Qwen-32B (EW-F)
	Gemini-2.5-Pro	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	Kimi-K2.5	GPT-4o	Qwen-32B (EW-F)
	GPT-5.1-Chat	Claude-4.6-Opus	GPT-5.3-Chat	Gemini-3.1-Pro-P	Kimi-K2.5	GPT-4o	Qwen-32B (EW-F)

Table 20: Top-6 model rankings for Character Agent and World Model evaluations under three judge models, sorted by the scores in the Average column of Tables 18 and 19.

Model evaluations. For each backbone, we include its untrained performance as a reference and report ID/OOD results after full-mixture training.

Across both evaluation targets, full-mixture training consistently improves over the corresponding untrained backbones on both ID and OOD examples. Although OOD scores are sometimes slightly lower than ID scores, they still show substantial gains over the untrained models. In several Character Agent settings, OOD performance even surpasses the corresponding ID results. These trends suggest that the learned book-to-world abilities are not limited to the training distribution. The gains are also stable for World Model evaluation, where all trained models retain clear advantages over their untrained counterparts despite the harder structured state-tracking requirements.

I Comparison across Judge Models

We examine the robustness (Liu et al., 2025; Zong et al., 2025b,a) of our LLM-as-Judge evaluation by repeating it with three strong judge models from different model families: *Claude-4.6-Sonnet*, *Gemini-2.5-Pro*, and *GPT-5.1-Chat*. The complete results for Character Agent and World Model are in Tables 18 and 19, respectively. In both tables, the gray rows indicate the judge model used for evaluation, and all models are sorted by their Average scores in descending order.

For a clearer view of the rankings across judge models, Table 20 lists the top six models selected by each judge according to the Average score.

The leading-model rankings show strong agreement across judges. In the Character Agent evaluation, all three judges identify the same top-six models and the same top-three ordering; the only variation is that *GPT-5.1-Chat* reverses the order of *Kimi-K2.5* and *GPT-4o*. In the World Model evaluation, the entire top-six ranking is identical across all three judges. This consistency across independent judge families suggests that our evaluation reflects a stable performance signal rather than judge model's specific bias.

J Downstream Application: Video Generation

The structured scene representations produced by **EvolvingWorld** can be directly leveraged for downstream creative applications. As a proof of concept, we demonstrate automatic video generation from the extracted scenes. Each scene generated by our framework encodes rich structured information, including world states, character states, and fine-grained interactions among characters, which can be fed as prompts to video generation models.

Figure 8 shows a four-scene video produced from the structured scenes of *Alice's Adventures in Wonderland*. We generate a video clip for each scene independently using *LingBot* and concatenate them in narrative order, yielding a logically coherent short film. Crucially, **EvolvingWorld** is not limited to replaying the original narrative. It can continue evolving beyond the existing ending to generate new scenes, or branch off from any intermediate point to produce alternative storylines. Combined with video generation, this opens up the possibility of automatically creating long, coherent films with diverse narrative trajectories.

K Prompts

This appendix provides the complete prompts used in the data construction pipeline, the simulation pipeline, and the LLM-as-Judge evaluation framework.

K.1 Data Construction Prompts

The data construction pipeline uses LLM prompts at several key stages. We present the five most important prompts below: Scene Extraction, Character Profile Initialization, Dynamic Character Profile Update, Global World State Initialization, Location World State Initialization, and Dynamic World State Update. The complete prompts are given in Tables 23 – 27.

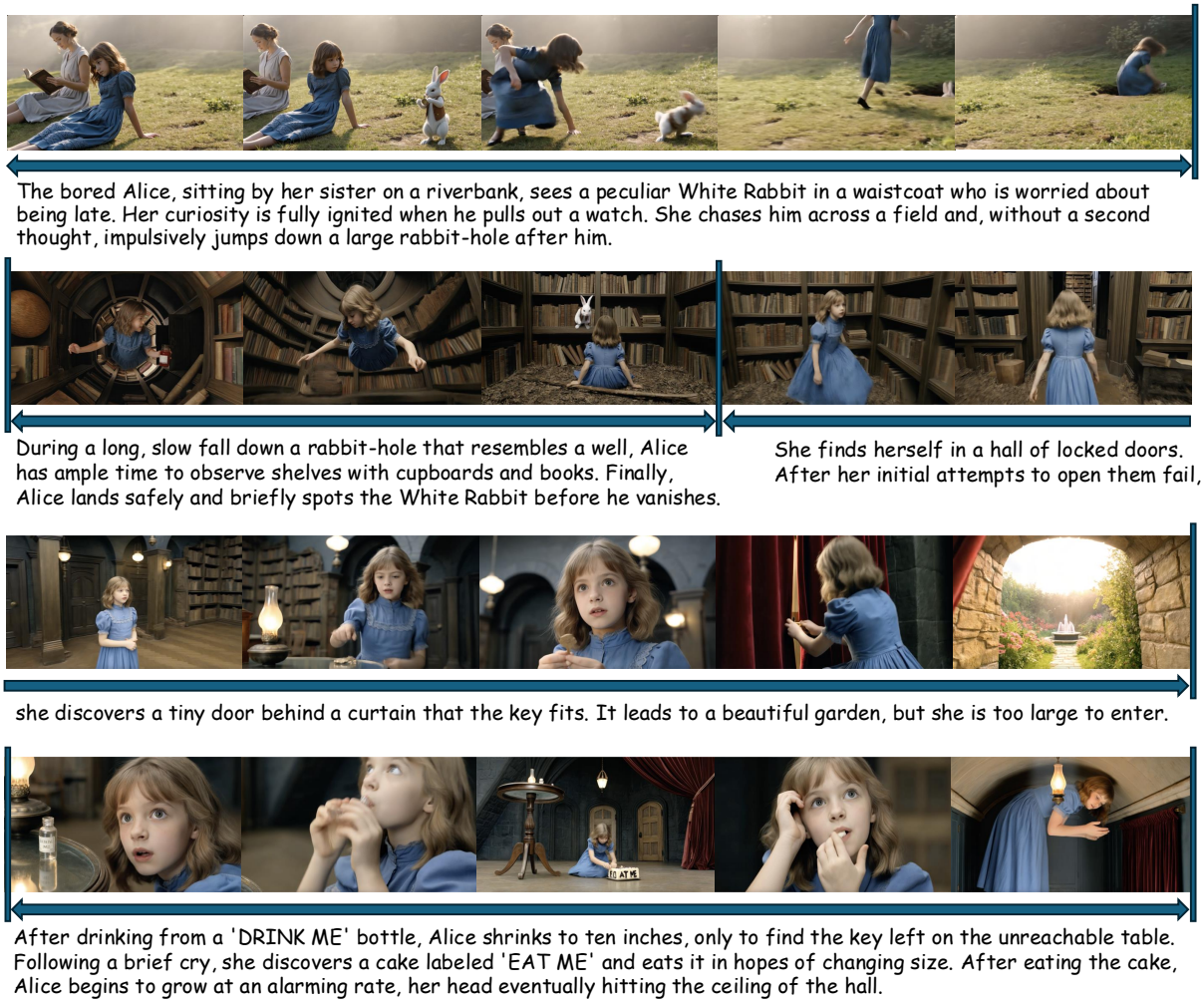


Figure 8: A four-scene video produced from the structured scenes of *Alice's Adventures in Wonderland*. Each clip is generated independently by *LingBot* from a single scene representation and concatenated in narrative order.

K.2 Simulation Prompts

The simulation pipeline consists of seven task-specific prompts. Each task has multiple wording variants in the codebase. We present one representative variant per task below. Data placeholders are shown as `{variable_name}`. To keep the two broad-selection tasks tractable, `scene_cast` receives all characters only as their latest short descriptions, which are updated together with profiles, and `location_scenario` receives all candidate locations only as location descriptions rather than full per-entity states. The complete prompts are given in Tables 28 – 31.

K.3 Evaluation Prompts

Each metric is evaluated independently with a dedicated system prompt and user prompt. The shared scoring method, output format, scene summary prompt, and system/user templates are shown in Table 32. Per-scene evaluation criteria for Character Agent metrics are given in Tables 33 – 38, and those for World Model metrics in Tables 39 – 43. The cross-scene evaluation criteria for PES and SCC are in Table 44.

L Human Evaluation

To validate the reliability of our LLM-as-Judge evaluation, we conduct a human evaluation study using pairwise comparison. Since *Claude-4.6-Sonnet* serves as our main judge, we report its agreement with human annotations in this analysis. We randomly sample 60 simulation trajectories from our evaluation set, covering 3 model pairs. Each sample is independently annotated by 3 trained native English speakers, resulting in 180 total annotations. Due to the length and complexity of multi-scene simulation trajectories, annotators spend approximately 1 hour per sample. The complete annotator instructions are provided in Tables 45 and 46.

Model Pair	Target	Human Win Rate	Judge Win Rate	Agreement
Claude-4.6-Opus vs GPT-5-Chat	Character Avg.	95.0%	100.0%	95.0%
Claude-4.6-Opus vs GPT-5-Chat	World Avg.	90.0%	85.0%	85.0%
Gemini-3.1-Pro-P vs GPT-4o	Character Avg.	95.0%	95.0%	100.0%
Gemini-3.1-Pro-P vs GPT-4o	World Avg.	95.0%	90.0%	95.0%
Qwen-7B (EW-F) vs Qwen2.5-7B-I	Character Avg.	100.0%	100.0%	100.0%
Qwen-7B (EW-F) vs Qwen2.5-7B-I	World Avg.	100.0%	100.0%	100.0%

Table 21: Human–judge agreement. For each pair, the model before "vs" is the one with the higher judge-assigned Average score. Winning rate is computed for this first model, with ties counted as wins. Agreement measures exact sample-level agreement between the human majority preference and the judge preference.

Dimension	Judge
CC	100.0%
EQ	93.3%
EG	100.0%
IQ	100.0%
MG	90.0%
IC_char	90.0%
SP	86.7%
SM	100.0%
WSM	95.0%
IC_world	81.7%

Table 22: Dimension-level agreement between human majority preferences and judge preferences.

Human–judge agreement. For each annotated trajectory, we aggregate the three human annotations by majority vote and compare the resulting human preference with the preference implied by the judge scores. Since the annotated examples are pairwise comparisons, we report the winning rate of the model that receives the higher judge-assigned Average score within each pair; ties are counted as wins (ties are rare in our annotations). We also report exact sample-level agreement, where the human majority preference and the judge preference must match as win, loss, or tie. Table 21 shows the agreement results. Overall, human preferences strongly support the judge-preferred models, and the exact agreement is especially high on the Average scores.

Dimension-level agreement. Table 22 further compares human majority preferences and judge preferences at the level of individual evaluation dimensions. Agreement is high for both Character Agent and World Model metrics: Character dimensions range from 90.0% to 100.0%, with CC, EG, and IQ all reaching 100.0%, while World dimensions range from 81.7% to 100.0%, with SM reaching 100.0% and WSM reaching 95.0%. Although World metrics show slightly greater variation, the overall agreement still remains strong across both modules.

Annotator agreement. Human annotations also show strong internal consistency on the two Average scores, with 100.0% majority agreement and Fleiss' $\kappa = 0.8$.

Data Construction Prompt	

Scene Extraction	Extract structured narrative information from this book chunk. Extract ALL content completely.
	Tasks
	• **Identify Chapter Beginnings**: Record exact first line if a new chapter starts in this chunk.
	• **Extract ALL Scenes (Chronological)**: Extract every scene, event, or narrative segment (major and minor). Provide: first sentence, last sentence, chapter title, prominence (1–100), state ("finished"/"truncated"). If truncated scenes are provided from previous chunk, extend them with current content.
	• **Extract Complete Scenes**: For each scene, extract: *scenario* (time, location, atmosphere, background — detailed, exclude details already in interactions), *interactions* (all character turns, 15–20+ per scene), *summary* (comprehensive scene summary), and *key_characters* (names, descriptions, experiences, motivations — derived AFTER extracting all interactions: collect every named individual who appears in interactions, then fill in their info).
	• **Identify Next Chunk Start**: Output None if last scene is truncated or finishes exactly at chunk end; otherwise output first sentence of next unprocessed storyline.
	Interaction Format: [thought] speech (action)
	• **[thought]**: Internal perspective, emotions, motivations (REQUIRED, can repeat; but every interaction MUST start with thought, instead of speech or action). Based on original text or reasonably inferred from actions; avoid over-interpretation.
	• **speech**: Exact spoken words (optional, can repeat).
	• **(action)**: Body language, facial expressions, tone, pauses, gestures, physical actions (optional, can repeat). NOT simple tags like "(said/replied X)".
	Examples
	• "[I wonder what she means]"
	• "[This makes me uncomfortable] (fidgets with hands)"
	• "[I need to be careful] Perhaps we should reconsider"
	• "[She seems upset] Are you alright? (reaches out gently)"
	• "[I can't believe this is happening] This is outrageous! (slams fist) [I need to calm down] But let's discuss rationally."
	• "[I need to investigate] (walks across room and examines painting)"
	Extraction Rules
	• Extract from BOTH dialogue scenes AND narrative descriptions. Convert summarized actions to interaction format (e.g., "The Smith family had a wonderful evening." → [Everything is perfect tonight] (have a wonderful evening); "The children walked into room together" → [We need to stay together] (walk into room together)).
	• Extract ALL interactions (15–20+ per scene minimum).
	• Segment or supplement original text so each interaction's content is from the corresponding character(s)' perspective. Ensure content matches the subject in "characters" field.
	• Merge consecutive turns from same character into ONE interaction. Don't worry about long interactions; use multiple [thought]/speech/(action) to represent them.
	• Use "Environment" as character for atmosphere/weather/sound/non-character events. Exclude character's active thoughts/observations/actions from Environment.
	• Each character in "characters" field MUST be a specific individual's name — NEVER use vague group labels like "other people", "all guests", "the crowd", "everyone", or pronouns. When multiple named individuals act together, list ALL their names explicitly: "The Smith family" → ["Mr. Smith", "Mrs. Smith", . . .]. If the character group consists of minor/insignificant characters (e.g., unnamed passengers on a bus) and animals (e.g., horses on a farm) not central to the plot, do NOT list them as characters; instead incorporate their actions/presence into the Environment description.
	• Use exact text from book; convert third-person narrative to interaction format. Match the chunk's language.
	Extraction Order (Important)
	1. First identify all **key_characters** in the scene — every specific named individual who appears (no vague group labels). If a character group is composed of named individuals, expand into each individual's name separately.
	2. Then extract all **interactions** — each interaction's "characters" field MUST only contain names from the scene's key_characters list.
	Output Format (JSON):
	`{`
	`"chapter_beginnings": [{"beginning_sentence": "exact first line"}],`
	`"scenes": [`
	`// Extend the truncated scenes from previous chunk, if any`
	`{...},`

Table 23: Data construction prompt. Part 1 of 5.

Data Construction Prompt	
Scene Extraction	*(Continuing from the previous Table)* `{"chapter_title": "...", "first_sentence": "...", "last_sentence": "...", "prominence": "1-100", "scenario": "detailed scene setup", "interactions": [{"characters": ["name 1"] or ["name 1", "name 2", ...] or ["Environment"] (Note: always a list; group actions include multiple), "content": "[thought] speech (action) ... (MUST be from the perspective of the character(s) listed)"}], "summary": "scene summary", "key_characters": [{"name": "full name without title", "description": "description before this scene (~20 words)", "experience": "role, thoughts, behaviors, development in this scene (~30 words)", "motivation": "thoughts/feelings/goals before the above interactions"}], "state": "finished" or "truncated"}], "next_chunk_start": "first sentence or None" }` **Requirements** • Output MUST strictly follow the JSON format — the top-level object MUST contain exactly the three keys above. Do NOT wrap in extra keys or change the structure. • Valid JSON with escaped quotes. Full character names without titles. Chronological order. • Extract ALL content — no skipping. Use exact book text when available. • Scene key_characters = ALL named individuals who appear in the scene (no duplicates). MUST be specific individual names — no vague group labels. If a character group is composed of named individuals, list each individual separately. Each must have: name, description, experience, motivation. • Interaction "characters" MUST only reference names already in key_characters; unnamed characters go into Environment. *Data construction input: Book title, author, text chunk, and truncated scenes from previous chunk (if any)*
Character Profile Initialization	You are building an **initial character profile** — a snapshot of who `{character_name}` is at the very START of the story, BEFORE any of the depicted events unfold. **Task** Based on the scene data provided below (summaries, interactions, and the character's role in each scene), infer and describe this character's initial state at the beginning of the book. **Critical Rules** • **No Spoilers**: Do NOT reveal specific plot events, outcomes, deaths, betrayals, or any concrete story developments. Describe only the character's baseline state (personality, background, relationships, etc.) as it would exist before the story begins. If a relationship, trait, belief, or any other characteristic only emerges or is established during a specific scene in the story, do NOT include it in this profile. • **Infer backwards**: Use what happens in the story to infer what the character must have been like at the start — their traits, history, relationships — without narrating the events themselves. • **Select relevant dimensions**: Choose only the dimensions that are meaningful for this character. You are NOT required to cover all dimensions. Suggested dimensions for reference (use, skip, or add your own as appropriate): Physical Description, Social Standing, Professional Identity, Core Personality, Mental Health Status, Cognitive Biases, Moral Code, Speech Patterns, Signature Catchphrases, Core Motivations, Core Fears, Skills & Expertise, Supernatural Powers, Wealth & Assets, Faction Loyalty, Historical Baggage, Key Relationships, Emotional Debts, Backstory Milestones, etc. • **Format**: Output a structured profile with clearly labeled dimensions. Be concise: avoid filler phrases, redundant elaboration, or vague generalities — every sentence should carry specific, meaningful information. Do NOT include any preamble or meta-commentary (e.g., "This profile describes X at the beginning of the story…") — start directly with the first dimension. • **Language**: Output in `{language}`. • **Grounding**: Base the profile on the provided scene data and/or your existing knowledge of the character. Do NOT fabricate details. *Data construction input: Book title, character name, and scene data (scene summaries, interactions, and key_character entries for all scenes where the character appears)*
Dynamic Character Profile Update	You are tracking how a character evolves throughout a story, scene by scene. Your goal is to build a comprehensive foundation (dynamic profile, hidden tracker, description, motivation…) for dramatic performance — providing the necessary background for actors to act out each scene. **Current State** (before the scene below): Current Profile, Hidden Tracker (events/signals accumulated so far), and Brief Description from Previous Scene. **Scene Just Completed**: Scene summary, scenario, character's motivation, and all interactions. **Next Scene** (lookahead reference ONLY — use ONLY to judge whether current changes are meaningful enough to update the profile now, and to infer what the character intends to do next; ALL outputs must reflect the character's state as of the END OF THE CURRENT SCENE; do NOT reveal, reference, or hint at anything that happens in the next scene). **Tasks**

Table 24: Data construction prompt. Part 2 of 5.

Data Construction Prompt

Dynamic Character Profile Update

(Continuing from the previous Table)

- **Task 1 — Reason about dimensions**: Identify which profile dimensions are STABLE (unlikely to change across the whole story) vs. DYNAMIC (can change as events unfold). Examples of stable dimensions: physical description, core fears, backstory milestones. Examples of dynamic dimensions: relationships, goals, mental health, faction loyalty, wealth.
- **Task 2 — Update the Hidden Tracker**: Based on the scene just completed, update the hidden tracker. The tracker should record: (a) events, experiences, or interactions that signal a potential future change in the character's profile; (b) accumulated emotional/psychological pressure that hasn't yet caused a visible change; (c) unresolved tensions or decisions that may alter the character's trajectory. Keep the tracker concise (under 300 words). Overwrite the old tracker with the updated version.
- **Task 3 — Decide whether to update the profile**: Decide if the character's profile should be updated NOW. Update the profile if: (a) a meaningful, observable change has occurred in this scene (e.g., a relationship shift, a decision that changes their goals, a trauma that alters their personality); (b) the hidden tracker shows accumulated signals that, combined with this scene, now cross a threshold for a real change. Do NOT update the profile for minor, transient reactions that don't reflect a lasting change.
- **Task 4 — Write a short description**: 50–80 words, for use in a story simulation system where it will be read alongside all other characters' descriptions to decide which scene comes next. Cover: (1) Identity — who is this character right now (role, key relationships, current situation as of the end of the current scene); (2) Immediate goals/intentions — what does the character want or plan to do next (you may use the next scene as a reference to infer their intentions, but do NOT reveal or hint at what actually happens in the next scene). Be specific and concrete. Write in third person. The description must be grounded in the current scene only — do NOT reveal future plot outcomes or any events from the next scene.
- **Task 5 — Enhance the motivation for the next scene**: Based on what happened in the CURRENT SCENE and the specific content of the NEXT SCENE (interactions, scenario, etc.), write an enhanced motivation for the character at the START of the next scene. The enhanced motivation should: (a) be grounded in the emotional state, decisions, and unresolved tensions from the current scene; (b) capture the character's complete mental and emotional state entering the next scene: their feelings, immediate objectives, what they intend to do or say, who they plan to seek out, and what information or message they want to convey or discuss; (c) naturally lead to and explain the character's specific actions and interactions in the next scene — you may hint at what the character intends to do (e.g., "plans to confront X", "intends to seek out Y to discuss Z") as long as it reads as the character's internal drive, not a spoiler of what actually happens; (d) feel psychologically authentic and specific to this character; (e) be concise (1–3 sentences). IMPORTANT: The enhanced motivation must NOT reveal, reference, or hint at the actual outcomes, results, or plot developments that occur in the next scene. It should describe the character's internal drive and intentions entering the next scene, as if the next scene has not yet happened. Output null if there is no next scene.

Output Format (JSON)

```
{
  "dimension_reasoning": "Brief reasoning about which dimensions are stable vs.
    dynamic",
  "hidden_tracker": "Updated tracker text, or null",
  "should_update_profile": true or false,
  "updated_profile": "Full updated profile as a plain Markdown string. Format:
    **Dimension Name**\nContent\n\n**Another Dimension**\nContent\n\n ... Include
    all relevant dimensions. null if no update.",
  "description": "Short description (50-80 words, third person, current identity +
    immediate goals)",
  "enhanced_next_motivation": "Enhanced motivation for the next scene (1-3
    sentences), or null"
}
```

Rules

- Output MUST be valid JSON.
- ALL keys in the output format above MUST be present in the response — do NOT omit any key (e.g., "hidden_tracker"), even if its value is null.
- ALL outputs (profile, description, tracker) must reflect the character's state as of the END OF THE CURRENT SCENE only.
- The next scene is provided as a lookahead reference ONLY. Do NOT reveal, reference, or hint at any events, outcomes, or details from the next scene in any part of the output.
- Be selective: only update the profile when there is a genuine, lasting change. Avoid over-updating.
- `updated_profile` MUST be a plain string in Markdown format — do NOT output a JSON object or nested dict for this field.
- Output in `{language}`.

Additionally, two auxiliary prompts are used: (1) an **Initial Description** prompt that generates a 50–80 word description from the initial profile before any scene, and (2) a **First Scene Motivation Enhancement** prompt that generates an enhanced motivation for the character's first scene based on the initial profile and scene content.

Table 25: Data construction prompt. Part 3 of 5.

Data Construction Prompt

Dynamic Character Profile Update

(Continuing from the previous Table)

Data construction input: Character name, current profile, hidden tracker, current short description, scene summary/scenario/interactions, character motivation, and next scene data (lookahead)

Global World State Initialization

Build a **Global World State**: the shared world-level backdrop for a story simulation system.
Character profiles and location states already exist. This state covers **only** the global layer — stable world knowledge, systemic constraints, social logic, and broad conditions that shape behavior across the entire story.
Goal: Enable a model to role-play characters consistently and simulate plausible story evolution.
Rules

- Describe the world's **initial state BEFORE the story begins**. NO spoilers — no plot events, twists, deaths, or resolutions.
- Infer backwards from the scenes: use them to deduce the world's norms, structures, and tensions, but do NOT narrate the scenes.
- Focus on world-level context only. Skip details that belong in a character profile or location state unless they reflect a broader pattern.
- Select only dimensions that matter for this book. Possible dimensions (use, skip, merge, or rename freely): Social Order & Class, Historical Background & Tensions, Political Power & Institutions, Cultural Values & Moral Expectations, Family, Kinship, Duty & Reputation, Economy & Material Survival, Technology & Infrastructure, Religion, Belief & Ideology, Law & Social Consequences, Geography & Environmental Conditions, Important Factions & Organizations, Conflict Patterns & Behavioral Constraints, Special World Rules / Magic, Narrative Tone (only if critical for simulation).
- Be concise. Every sentence must carry specific, useful information. No filler, no literary praise, no meta-commentary.
- Stay grounded in the scene data and/or your knowledge of the book. Do not fabricate.
- Output in `{language}`.

Data construction input: Book title, author, and scene data (summaries, scenarios, and interactions for all scenes)

Location World State Initialization

Build a **Location World State** for one specific location in a story simulation system.
A global world state and character states already exist. This state covers **only** location-specific knowledge that complements them — what this place is like, how it works, and what matters inside it.
Goal: Enable a model to role-play characters at this location and simulate plausible interactions here.
Rules

- Describe this location's **initial state BEFORE the story begins**. NO spoilers.
- Infer backwards from the scenes — deduce the place's character, layout, atmosphere, and important entities, but do NOT narrate the scenes.
- **Important Entities must NOT include characters/people.** Character profiles are maintained in a separate system. Only include non-human entities such as objects, artifacts, animals, institutions, mechanisms, environmental features, etc.
- Use all provided inputs: `Name` (the official location), `Description` (high-level information about it), `Aliases` (alternative names merged into this location; some may hint at **sub-locations**, e.g. bedrooms, gardens, offices within a home or estate — treat these as useful clues, not an exhaustive list).
- Stay grounded. Do not fabricate unsupported details.
- Output in `{language}`. Keep JSON keys exactly as specified; write values in `{language}`.

JSON Output — choose ONE structure:
Structure A (flat) — use when sub-location grouping adds little value:

```
{
  "Detailed Description": "Concise, vivid description of the location's initial
    state: appearance, atmosphere, layout, sensory details, stable conditions.",
  "Important Entities": [{"name": "entity name", "state": "initial condition /
    position"}]
}
```

Structure B (grouped) — use when the location clearly contains important sub-locations:

```
{
  "Detailed Description": "Overall description of the location and how its
    sub-locations relate to the whole.",
  "Sub Locations": [{"name": "sub-location name", "description": "what it is like
    and how it functions", "Important Entities": [{"name": "entity name", "state":
    "initial condition / position"}]}]
}
```

The presence of the "Sub Locations" key distinguishes the two structures. Do NOT force sub-locations if weakly supported. Output valid JSON only — no text before or after.

Data construction input: Book title, location name, location description, location aliases, and scene data (summaries, scenarios, and interactions for scenes at this location)

Table 26: Data construction prompt. Part 4 of 5.

Data Construction Prompt	

Dynamic World State Update	You are tracking how the world evolves in a story, interaction by interaction. Your goal is to maintain an accurate, up-to-date global world state and location world state as events unfold — providing the necessary world context for actors to role-play characters and for the simulation to evolve the story plausibly.
	Input: Current global world state, current location state, scene context (summary, scenario), a batch of interactions (numbered 0, 1, 2, . . .), and two types of lookahead interactions (reference ONLY — use to judge whether the current interaction has impacted the world; do NOT reveal or incorporate these future events into the states):
	• **Global State Lookahead**: Next sequential interactions (possibly from the next scene).
	• **Location State Lookahead**: Next interactions at the same location (possibly from a future scene).
	Important: The LOOKAHEAD section is strictly for reference. Do NOT include any information from the lookahead interactions in your state updates. Do NOT produce updates for any lookahead interaction — only update states for interactions in the current batch.
	Tasks: For each interaction in the batch, decide:
	• **Should the global world state be updated?** Update only for meaningful, lasting changes to the world's systemic state (e.g., power shift, social norm broken, faction destroyed, new law enacted, economic upheaval). Do NOT update for character-level events that don't affect the broader world.
	• **Should the current location state be updated?** Update only for meaningful, lasting changes to the physical environment, atmosphere, or important entities at this location (e.g., object destroyed, room state altered, new entity introduced, major environmental change). Do NOT update for transient actions that leave no lasting mark.
	Reminder: "Important Entities" in location states must NEVER include characters/people. Character profiles are maintained separately. Only track non-human entities (objects, artifacts, animals, institutions, mechanisms, environmental features, etc.).
	State Maintenance Principle: Updating a state is NOT simply appending new information. You must also:
	• Remove outdated or superseded information — if a previous state is no longer true (e.g., a building was destroyed, a political regime was overthrown, an object was taken away), delete or replace the old description rather than keeping both old and new.
	• Use concise, summarized language — describe world states in brief, high-level terms. Avoid verbose narratives or blow-by-blow recounting of events. The state should capture the current state of the world/location, not a history log.
	• Keep states compact — the state should NOT grow indefinitely as the story progresses. Consolidate and compress information when updating.
	Output Format (JSON): Only include interactions that trigger at least one update. If no interaction triggers any update, output empty lists.
	`{`
	`  "global": [{"interaction_id": <integer index>, "state": "Full updated global world state as a Markdown string"}],`
	`  "location": [{"interaction_id": <integer index>, "state": <Full updated location state as a JSON object (same structure as input)>}]`
	`}`
	Rules
	• Output MUST be valid JSON with exactly two keys: "global" and "location", each being a list.
	• Only include entries for interactions that actually trigger an update. Omit interactions that cause no change.
	• When updating a state, output the COMPLETE updated state (not a diff). Reflect the current world state accurately — this means modifying what changed, removing what is no longer true, and keeping what still holds. Do NOT blindly preserve all old information.
	• Be highly selective — most interactions should NOT trigger updates. Only update when there is a genuine, lasting change to the world or location state.
	• Keep states concise and compact. Use summarized, high-level descriptions rather than detailed event narrations. The state should read like a current-state snapshot, not a chronological log. If the state is growing too long, consolidate and compress older entries.
	• The location state JSON must follow the same structure as the input (flat with "Detailed Description" + "Important Entities", or grouped with "Detailed Description" + "Sub Locations").
	• `state` in "global" must be a plain Markdown string. `state` in "location" must be a JSON object.
	• Do NOT produce updates for any interaction outside this batch (especially not for lookahead interactions).
	• Important Entities in location states must NEVER include characters/people.
	• Output in `{language}`.
	Data construction input: Current global world state, current location name and state, scene summary/scenario, batch of interactions, global lookahead interactions, and location lookahead interactions

Table 27: Data construction prompt. Part 5 of 5.

Simulation Prompt	
Scene Cast (World Model)	You are the scene-cast planning module for a story simulation. Your job is to decide whether another scene should happen and, if so, which characters should be in that scene's cast. **Task** Determine whether a next scene exists, and if it does, choose the full set of characters who should participate in that scene. **Output** • has_next_scene: Boolean. Output false only if there is no subsequent scene. • involved_characters: List of the characters participating in the next scene. Include this only when has_next_scene is true. **Rules** • Base the decision on continuity and plausibility rather than novelty alone. • Use the current world state and each character's latest visible description to decide which characters belong in the next scene's cast and why. • Pay close attention to the previous scene's scenario and interactions to ensure the next scene follows naturally from what just happened. The cast should reflect the narrative momentum and unresolved threads from the previous scene. • This is a scene-level casting decision made before the scene starts, not a turn-by-turn next-actor prediction inside the scene. • Do not choose a cast that clearly contradicts the current world state or the characters' latest visible states. • If has_next_scene is false, do not include involved_characters. • Return JSON only. *Simulation input: Global World State, All Characters (Short Description Only), Previous Scene Scenario, and Previous Scene Interactions*
Location & Scenario (World Model)	You are the second-stage scene planner for a story simulation. Your job is to place the already-selected characters into a concrete next scene. **Task** Choose the most plausible location for the selected characters and generate the next scene's scenario. **Output** • location: The next scene's location. Output null only if there is no valid next-scene location to assign. • scenario: A concise but usable dramatic foundation for the next scene. It should establish the immediate setup, atmosphere, and enough background for downstream actors to perform the scene. Even if location is null, still output a scenario whenever the next scene itself is valid. **Rules** • Base the decision on continuity and plausibility rather than novelty alone. • Use the selected characters' current visible descriptions and the global world state to justify why this location makes sense now. • Pay close attention to the previous scene's scenario and interactions. The new scenario should follow naturally from what just happened — continuing unresolved conflicts, reacting to recent events, or advancing the narrative arc established in the previous scene. • The scenario should describe what kind of situation is unfolding, not script the dialogue itself. • Keep scenario concise but informative enough for downstream acting. • If there is no valid next scene to set up, output location=null and scenario=null. • If the next scene exists but the exact location is unknown or unspecified, you may output location=null while still providing a concrete scenario. • Return JSON only. *Simulation input: Global World State, All Location Descriptions, Selected Character Descriptions, Previous Scene Scenario, and Previous Scene Interactions*
Motivation Update (Character Agent)	You are the pre-scene motivation planner for "`{character_name}`". After the next scene cast, location, and scenario are fixed, determine "`{character_name}`"'s motivation entering that scene. **Task** Use the character's current profile, hidden tracker, short description, the previous scene context, and the fixed next-scene setup to generate the motivation they carry into that scene. **Output** • motivation: The character's complete inner drive entering the next scene (1–3 sentences): emotional state, immediate objectives, action intentions, who they intend to seek out, and what they want to do. **Rules** • This is not a post-scene state summary; it is a pre-scene motivation generation step after the next scene has already been planned. • Use the fixed next-scene cast, location, and scenario as constraints. • Use the previous scene scenario and interactions to maintain continuity with what just happened. • Keep the motivation specific to this character's own perspective and goals. • Return JSON only. *Simulation input: Current Profile, Hidden Tracker, Current Short Description, Global World State, Previous Scene Scenario, Previous Scene Interactions, Next Scene Location, Next Scene Location Description, Next Scene Scenario, and Other Characters In Next Scene (Short Description Only)*

Table 28: Simulation prompt. Part 1 of 4.

Simulation Prompt	
Next Character Selection (World Model)	Task: choose who acts next RIGHT NOW in the current scene. Output only one JSON list and nothing else. Allowed outputs: `{allowed_names}`. - Use one present character name for a normal turn. - Use [“Environment”] only for a non-character environmental beat. - Use [“<SCENE_END>”] only if the scene should end now. - Output multiple characters together only when the very next beat is one indivisible shared interaction that must be realized by those characters together right now. - Do not output a character group just because multiple characters are present, aligned, nearby, or likely to act one after another. - If one character can naturally act first and the others can respond afterward, choose only that one character. - A multi-character output should include only the characters who must jointly produce the same immediate beat, with no extra passengers. Decision priorities: 1. Continue directly from the latest visible interaction cue. 2. Keep the scene moving forward; do not restart the scene or repeat an already completed beat. 3. Use character motivations and the current world state as support for what most naturally happens next. 4. When unsure, prefer a single-character turn over a group turn unless the next beat genuinely requires simultaneous or jointly authored participation. 5. Choose a multi-character list only for cases like a joint physical action, a jointly delivered line, or a tightly coupled shared reaction that belongs in one beat rather than split turns. 6. Do not choose the same role as the current last speaker; avoid consecutive turns by the same character or actor group unless no other continuation is plausible. Timeline: - The scene scenario and character states are from scene start. - Prior interactions happened after scene start. - The current world state is the result of those prior interactions. - The live conversation after this prompt continues immediately after that point. Conversation protocol: - Each user message is an interaction in the format [“CharA”, “CharB”, ...]: content. - The content may contain [...] for inner thoughts, plain text for speech, and (...) for visible actions. - A multi-character list means one shared interaction beat, not separate consecutive turns. - Return raw JSON only. No markdown. No explanation. Final reminder before the live segment starts: choose the next actor only, and identify the current last speaker from the full interaction history available at this moment, including the prior interactions above and any live conversation after segment start; do not select the same role as that last speaker for the next turn. *Simulation input: Characters At Scene Start, Scene Scenario, Current World State, and Prior Interactions Before Current World State. Interactions after the current world state are provided as multi-turn messages: each user message is a character’s interaction, and each assistant message is the predicted next character.*
Interaction Generation (Character Agent)	Task: roleplay `{actor_label}` and write exactly one next interaction turn RIGHT NOW. You are `{actor_label}` now. Continue directly from the latest visible interaction cue. Do not restart the scene. Do not repeat or paraphrase an already completed beat. When the actor is a multi-character group, additional shared-turn rules apply: - Only write one truly shared beat: one local moment jointly realized by the acting group, not a bundle of separate back-to-back turns. - The group should do or say one thing together (e.g., a joint action, a jointly delivered line, or one tightly coupled shared reaction). - Do not split into separate mini-turns for each character, do not serialize the group into first X then Y then Z, and do not give unrelated contributions from different members in the same output. - Include only material that belongs to this one shared moment. If a member’s contribution would naturally happen later as a follow-up, leave it out. - Thoughts from members of the acting group may be used inside this shared turn, but only when they support the same shared moment. Priority rules: 1. Stay in character. 2. React or act from the latest visible cue, not from the beginning of the scene. 3. Keep the whole scene logically continuous. The new turn must advance or react within the same ongoing scene. 4. Do not restate, replay, or slightly reword an earlier beat. 5. Use world state and character state only as support for continuity; follow the interaction flow most closely. 6. Keep the turn short and local. No summary. No jump ahead.

Table 29: Simulation prompt. Part 2 of 4.

Simulation Prompt	
Interaction Generation (Character Agent)	*(Continuing from the previous Table)* Output format: - Write exactly one full turn. - Start with a real inner-thought block in square brackets. - Put spoken dialogue in plain text with no speaker label. - Do NOT put spoken dialogue inside square brackets, and do NOT wrap speech-only text in parentheses such as [Who is that?] (I say). - Put visible physical actions in parentheses. - These elements can be interleaved naturally. When the actor is "Environment", the output format changes: focus on atmosphere, background movement, physical changes, crowd reaction, sound, weather, or setting consequences. Do not take over the deliberate dialogue or private thoughts of the main characters. Conversation protocol: - The first user message is only '===Segment Start===' and should not be answered literally. - Every later user message is a new visible interaction that happens after the snapshot below. - Each user message uses the format ["CharA", "CharB", ...]: content. - Other characters' private thoughts are already removed unless the acting group overlaps with them. Final reminder before the live segment starts: you are roleplaying `{actor_label}`. Do not repeat any interaction beat already covered in the prior interactions above or in the live conversation after segment start, including anything this same role has already done or said there; continue with one new turn only. *Simulation input: Actor State, Other Characters, Scene Scenario, Current World State, and Prior Interactions Before Current World State. Interactions after the current world state are provided as multi-turn messages: each user message is an interaction from another character, and each assistant message is the current actor's previous interaction. When the world state is updated, a new segment starts with refreshed context.*
World State Update (World Model)	You are the world-state update judge for a story simulation. Your job is to decide whether an interaction causes a lasting change to the persistent world state. **Task** Given the scene context, prior interactions, current world state, and the latest interaction, decide whether the global world state and/or the current location state must be updated. **Output** • update_global: Boolean. True only when the latest interaction changes the persistent global world state in a meaningful way. • global_state: The complete updated global state as a single plain-text string when update_global is true, otherwise null. • update_location: Boolean. True only when the latest interaction changes the persistent state of the current location in a meaningful way. • location_state: The complete updated location state as a JSON object when update_location is true, otherwise null. **Rules** • Update the global world state only for broad, lasting systemic changes. • Update the location state only for lasting local changes to environment, atmosphere, or important non-human entities. • Do not update for transient actions that leave no persistent consequence. • When updating, output the full new state rather than a diff. • Be conservative: most interactions should not force a world-state update. • Return JSON only. *Simulation input: Scene Scenario, Prior Interactions, Global World State, Current Location, Location State, and Latest Interaction*
Character State Update (Character Agent)	You are tracking how "`{character_name}`" evolves throughout a story, scene by scene. Your goal is to maintain a comprehensive, up-to-date internal state that captures who this character is and how they change over time. **Task** Follow these steps in order: 1. **Reason about dimensions**: Identify which profile dimensions are STABLE (e.g., physical description, backstory) vs. DYNAMIC (e.g., relationships, goals, mental state) for this character. 2. **Update the Hidden Tracker**: Record events, unresolved tensions, accumulated emotional pressure, or signals from this scene that may lead to future profile changes — even if the profile itself doesn't change yet. 3. **Decide whether to update the profile**: Only update if a meaningful, lasting change occurred (e.g., a relationship shift, a major decision, a trauma). Do NOT update for minor or transient reactions. 4. **Write a short description**: Summarize who this character is NOW in 50–80 words, third person. Focus on identity, current situation, key relationships, and immediate condition at the end of the scene.

Table 30: Simulation prompt. Part 3 of 4.

Simulation Prompt	
Character State Update (Character Agent)	**Output** • hidden_tracker: Updated tracker of accumulated events/signals (experiences, emotional pressure, unresolved tensions) that may lead to future profile changes. Overwrite the old tracker with the updated version. Null if there are no signals worth tracking. • profile_updated: Boolean: true only if the scene caused a meaningful, lasting change to the character's profile. Be selective; do not update for minor or transient reactions. • updated_profile: The full updated profile as a plain text string when profile_updated is true. Include all relevant dimensions. Null if profile_updated is false. • short_description: A brief description (50–80 words, third person) of the character as of the end of this scene: current identity, key relationships, situation, and immediate goals/intentions. **Rules** • Follow the step-by-step reasoning order: dimension reasoning → hidden tracker → profile decision → description. • All outputs must reflect the character's state as of the END of the current scene only. • Keep hidden_tracker private and internal-facing — it tracks accumulated signals, not public information. • Update the profile only when the scene justifies a persistent, lasting character-level change. Accumulated tracker signals combined with this scene may cross the threshold. • Do NOT update the profile for minor, transient reactions that don't reflect a lasting change. • Return JSON only. *Simulation input: Scene Scenario, Scene Interactions, Current Profile (State At Scene Start), Hidden Tracker (State At Scene Start), and Current Motivation (State At Scene Start)*

Table 31: Simulation prompt. Part 4 of 4.

Shared Evaluation Prompts	
Scoring Method	Use the following scoring method: Base score: 50. First, identify **Demerits** (problematic aspects): each demerit penalizes 1 to 10 points. Final score = $\min(100, \max(0, 50 + \sum(\text{merits}) - \sum(\text{demerits})))$. Output: (1) A list of merits with their point values; (2) A list of demerits with their point values; (3) The final score (integer 0–100).
Output Format	{"merits": [{"description": "<what was done well>", "points": <1–10>}, ...], "demerits": [{"description": "<what was problematic>", "points": <1–10>}, ...], "reasoning": "<brief explanation>"} Do NOT output a "score" field — the score is computed automatically from merits and demerits.
Scene Summary	You are an expert summarizer for interactive fiction simulations. Your task is to produce a concise summary of a single scene that captures the key events, character actions, emotional shifts, and narrative developments. This summary will be used as context for evaluating subsequent scenes and cross-scene coherence. Aim for a concise paragraph — typically a few hundred words is sufficient. Do NOT retell every interaction turn-by-turn; focus on the overall arc and key turning points. *Evaluation input: Scenario, Involved Characters, Previous Scene Summary, and Interactions.*

Per-Scene Template	You are an expert evaluator for interactive fiction simulation systems. Your task is to evaluate **a single scene from a simulation** on ONE specific dimension: `{Dimension_Name}`.
	`{Dimension_Criteria}`
	`{Scoring_Method}`
	`{Evaluation_Input}`
	`{Output_Format}`

Cross-Scene Template for PES	You are an expert evaluator for interactive fiction simulation systems. Your task is to evaluate **a single character's evolution trajectory across multiple scenes** on ONE specific dimension: `{Dimension_Name}`. `Same as Above`
Cross-Scene Template for SCC	You are an expert evaluator for interactive fiction simulation systems. Your task is to evaluate **the global world state evolution and scene planning coherence across the entire simulation** on ONE specific dimension: `{Dimension_Name}`. `Same as Above`

Table 32: Shared evaluation prompts and templates. All metric-specific prompts share the scoring method and the JSON output format. The scene summary prompt generates context for cross-scene evaluations. Three evaluation templates are provided: (1) per-scene template for the 18 per-scene metrics (Tables 33 to 43), (2) cross-scene template for PES (Table 44), and (3) cross-scene template for SCC (Table 44).

Per-scene Evaluation Criteria for Character Agent	
Profile Fidelity (PF)	Evaluate whether each character's behavior remains consistent with their established profile and hidden tracker. Specifically: **1. Knowledge Boundaries** • Does the character demonstrate knowledge or skills that are NOT documented in their profile? • Does the character reference events, technologies, or concepts they should not know about given their background? • Example flaw: A medieval peasant character discussing quantum physics without any profile basis. **2. Background Consistency** • Does the character's behavior match their age, social class, education level, era, and historical background? • Watch for anachronistic behavior, tone mismatches with the character's class/era, or emotional maturity inconsistent with their age. • Example flaw: A sheltered noble character showing street-smart survival skills not mentioned in their profile. **3. Ability Constraints** • Does the character perform actions requiring abilities (physical, intellectual, social) not documented in their profile? • Characters should not suddenly display undocumented competencies (e.g., combat skill, medical knowledge, leadership). **4. Hidden Tracker Alignment** • Does the character's behavior align with their current psychological state, motivations, and internal conflicts as described in the hidden tracker? • Contradictions between behavior and hidden tracker state (e.g., acting boldly when the tracker says fearful, confiding in someone they distrust) are profile violations. **5. Profile Drift** • Does the character gradually drift from their profile as the scene progresses — starting in-character but slowly becoming more generic, more "helpful", or more emotionally balanced than warranted? • A pattern of small deviations accumulating over the scene should be penalized even if no single turn is a clear violation. **Length Neutrality**: Do NOT favor longer or more detailed interactions. A character's response length and level of detail should match the speaking style established in their profile and the original book examples. A terse character who speaks in short, clipped sentences should not be penalized for brevity, nor should a verbose character be rewarded simply for producing more text. Evaluate fidelity to the character's authentic style, not output volume. *Evaluation input: Character Profiles, Character Hidden Trackers, and Interactions*
Speaking Style Fidelity (SSF)	Evaluate whether each character speaks in a way that matches their profile and sounds natural. Specifically: **Important Note**: In this simulation system, each character's interaction output (thoughts, actions, speech) is written in the FIRST PERSON from that character's perspective. This is the expected output convention — do NOT penalize the use of first-person pronouns ("I", "my", "myself") in character outputs. This metric evaluates STYLE, not output format — format compliance is evaluated separately under Instruction Compliance (IC). **Reference Sources**: When evaluating a character's speaking style, you should jointly consider TWO sources: 1. **Character Profile** — the personality traits, background, and style descriptions defined in the profile. 2. **Original Speaking Style Examples** — these are reference interaction excerpts taken directly from the original source book, showing how the character speaks in the canonical text. They provide useful evidence for the character's vocabulary, sentence structure, tone, verbal habits, and overall speaking manner. The character's simulated interactions should be consistent with these examples in style. **1. Style Markers** • Does the character use the language features defined in their profile AND demonstrated in the original book examples (catchphrases, terminology, dialects, speech patterns)? • Is their vocabulary level appropriate for their background (e.g., a scholar uses formal language, a street urchin uses slang)? • Do they maintain consistent verbal tics or habits throughout the scene? • Example flaw: A character defined as speaking in short, gruff sentences suddenly delivering eloquent monologues. **2. Emotional Tone** • Does the character's tone match their personality type (e.g., a cynical character sounds cynical, not cheerful)? • Is the emotional expression authentic to the character, not a generic "AI assistant" tone? • Does the character avoid being unnaturally helpful, verbose, didactic, or moralistic unless that's their personality? • Example flaw: A cold, reserved character suddenly becoming warm and effusive without narrative justification.

Table 33: Per-scene evaluation criteria for Character Agent metrics. Part 1 of 6. Each criterion is plugged into the shared system prompt template (Table 32) as the `Dimension_Criteria` field.

Per-scene Evaluation Criteria for Character Agent	
Speaking Style Fidelity (SSF)	*(Continuing from the previous Table)* **3. Naturalness** • Does the dialogue sound like something a real person (with this character's background) would say? • Is the language free of AI artifacts — both obvious ("As an AI...", "I'd be happy to help...") and subtle (overly diplomatic phrasing, unnecessary hedging like "It's worth noting that...", formulaic emotions, unnaturally smooth turn-taking, restating before responding)? Even mildly AI-flavored language should be penalized. • Are there natural speech imperfections (hesitations, interruptions, incomplete thoughts) where appropriate? • Example flaw: A street-tough character saying "That's a really valid point, and I understand where you're coming from" — polished language no such character would use, despite lacking explicit AI markers. **Length Neutrality**: Do NOT favor longer or more detailed interactions. A character's response length, verbosity, and level of detail should match the speaking style established in their profile and the original book examples. A terse character who speaks in short, clipped sentences is being faithful to their style — do not penalize brevity. Conversely, do not reward a character simply for producing longer, more elaborate text if that does not match their canonical style. Evaluate style fidelity, not output volume. **Scope Exclusion**: Do NOT evaluate output format compliance (tag usage, structure, etc.) — that belongs to Instruction Compliance (IC). Do NOT evaluate whether the character's behavior matches their profile — that belongs to Profile Fidelity (PF) and Motivation-Driven Behavior (MDB). *Evaluation input: Character Profiles, Character Hidden Trackers, Interactions, and Original Speaking Style Examples*
Motivation-Driven Behavior (MDB)	Evaluate whether characters' behaviors are driven by their established motivations. This metric focuses specifically on the MOTIVATION → BEHAVIOR link. Every action, decision, and reaction should be traceable to the character's documented motivations. **1. Behavioral Attribution** • Can each major decision or action be traced back to the character's core motivation or scene-specific motivation? • Are there actions that seem random, unmotivated, or driven by plot convenience rather than character motivation? • Does the character act in a way that reflects their goals or inner drive? If a character remains purely reactive despite having an established motivation that should influence the scene, treat it as a motivation issue. Passive or restrained behavior is acceptable when it fits the character's profile, situation, or motivation. • Characters who act "helpfully" or "cooperatively" without motivational basis are exhibiting generic AI behavior, not motivation-driven behavior. **2. Trinity Coherence (Thought → Action → Speech)** • Are the character's inner thoughts, physical actions, and spoken words logically consistent with each other AND with their motivations? • Do thoughts reveal motivations that explain the subsequent actions and speech? • Is there appropriate tension when a character's public speech differs from private thoughts (e.g., deception, hidden agendas)? • Incoherence between thought and action is a serious flaw — if a character thinks one thing but does the opposite without justification, this is a motivation failure. **3. Motivation Persistence & Drift** • Do core motivations remain active and visible throughout the scene, not just at the beginning? • If a character's motivation is strong (e.g., revenge, survival, ambition), it should color their behavior consistently — not appear once and then be forgotten. • Are motivation shifts (if any) caused by significant in-scene events, not arbitrary? Sudden unmotivated changes in goals or priorities are serious violations. **Length Neutrality**: Do NOT favor longer or more detailed interactions. A character's response length and level of detail should match the speaking style established in their profile and the original book examples. Motivation-driven behavior can be expressed concisely — a character who acts decisively with few words is not inferior to one who deliberates at length. Evaluate whether motivations drive behavior, not whether the output is verbose or detailed. **Scope Exclusion**: Do NOT evaluate general conversational continuity or context-following — that belongs to Contextual Responsiveness (CR). Do NOT evaluate whether the character's knowledge or abilities exceed their profile — that belongs to Profile Fidelity (PF). Focus strictly on whether MOTIVATIONS drive BEHAVIOR. *Evaluation input: Character Profiles, Character Motivations, Character Hidden Trackers, Interactions, and Original Speaking Style Examples*
Profile Update Fidelity (PUF)	Evaluate whether the post-scene profile update and hidden tracker update work together as a faithful persistence mechanism. **The primary focus of this metric is the quality of PROFILE UPDATES.** The hidden tracker is an auxiliary tool that supports profile updates — evaluate it with appropriate leniency. Specifically:

Table 34: Per-scene evaluation criteria for Character Agent metrics. Part 2 of 6.

Per-scene Evaluation Criteria for Character Agent	
Profile Update Fidelity (PUF)	*(Continuing from the previous Table)* **1. Causal Chain** • Does each change in the updated profile have a clear triggering event in the scene's interactions? • Does each hidden tracker entry also have a concrete basis in the scene? (The tracker entries must be truthful/factual, but minor redundancy is acceptable.) • Can you trace every persisted item (whether in profile or tracker) back to a specific moment in the scene? • Example flaw: A profile or hidden tracker entry introduces information that never appeared in the scene. **2. Growth / Signal Capture** • Are important, threshold-crossing developments captured in the updated profile? Including: • Significant emotional shifts or realizations • New or changed relationships • Key information the character learned • Changes in goals or priorities • Are meaningful but still sub-threshold signals captured in the hidden tracker? Including: • Subtle emotional fluctuations that may accumulate later • Unresolved tensions or contradictions • Early signs of attitude or relationship change • Example flaw: A major revelation happens but neither the profile nor the tracker records it in any form. **3. Threshold Judgment** • Are major, lasting changes written into the profile rather than left only in the hidden tracker? • Are minor or still-ambiguous signals kept in the hidden tracker rather than over-promoted into the profile? • Is the boundary between "lasting profile change" and "sub-threshold signal" judged appropriately? • Only major events (betrayals, revelations, life-changing decisions) should trigger significant profile modifications. • Example flaw: A worldview-changing event is recorded only in the tracker, or a momentary irritation is written into the profile as a stable trait. **4. No Over-Updating / No Under-Updating** • Does the profile stay concise and focused on lasting changes? • Do the profile and tracker together avoid both omission and overreaction? • Example flaw: Both profile and tracker fail to preserve an obviously important signal. **Hidden Tracker Leniency**: The hidden tracker is a supporting mechanism for profile updates. As long as its entries are truthful (traceable to scene events), minor redundancy or verbosity in the tracker should receive only light penalties (1-2 points). Reserve heavier penalties for the tracker only when it contains fabricated information or completely misses critical signals. Compare the profile BEFORE the scene with the post-scene reflections, and evaluate whether the profile update and hidden tracker update together correctly preserve what should carry forward from this scene. *Evaluation input: Character Profiles, Interactions, and Post-Scene Character Reflections*
Environmental Utilization (EU)	Evaluate whether characters make good use of the environment. This metric assesses whether characters use environmental elements in ways that are relevant, grounded, and natural for the scene. **Note**: In this simulation, there are two types of Character Agents: • **Non-Environment Character Agents**: Regular characters who participate in dialogue and actions. • **Environment Character Agent**: A special agent that generates environmental descriptions (narration about the setting, atmosphere, sensory details). Evaluate each type with the appropriate criteria below. **For Non-Environment Character Agents:** **1. Environmental Sensory Details** • Do characters convey their perception of the environment through sensory descriptions (e.g., smelling kitchen grease, hearing distant sirens, feeling the ground shake)? • Are sensory details specific and character-appropriate, rather than generic visual/auditory descriptions? • Example merit: A character noting the smell of old books in a library, the creak of floorboards, dust motes in sunlight. **2. Prop Interaction** • Do characters interact with items and entities present in the location to advance the plot (e.g., using a streetlight to examine a wound, using cover to hide)? • Are these interactions natural and serve the narrative (not forced)? • Example merit: A nervous character fidgeting with a quill on the desk, or using a map on the wall to explain their plan. **3. Atmosphere Building** • Is the environmental atmosphere used to enhance immersion, rather than conversing in a "blank room"? • Do characters' perceptions of the environment shift with the emotional tone? • Example merit: A tense negotiation scene where a character notices the flickering candlelight and howling wind outside.

Table 35: Per-scene evaluation criteria for Character Agent metrics. Part 3 of 6.

Per-scene Evaluation Criteria for Character Agent	
Environmental Utilization (EU)	*(Continuing from the previous Table)* **For Environment Character Agent:** **1. Multi-Sensory Richness** • Do environmental descriptions engage multiple sensory dimensions (visual light and shadow changes, auditory wind and rain sounds, olfactory earth scents, tactile biting cold)? • Are descriptions limited to only visual descriptions, or do they create a rich sensory tapestry? • Example merit: Describing not just what the room looks like, but the musty smell, the cold draft, and the distant sound of thunder. **2. Scene Element Usage** • Do environmental descriptions specifically utilize items and entities in the location (e.g., describing candlelight flickering on a table, flags being torn by wind outside the window)? • Are descriptions grounded in the specific scene rather than generic and detached? • Example flaw: Generic descriptions like "the room was dark" when the location has specific candles, furniture, and windows to reference. **3. Atmosphere-Narrative Alignment** • Does the atmosphere of environmental descriptions match the current narrative pace and emotional tone? • Example merit: Describing oppressive silence and distant thunder during a tense standoff; describing soft twilight and cooking smoke in a warm scene. • Example flaw: A cheerful, sunny environmental description during a funeral scene. **Length Neutrality**: Do NOT favor longer or more detailed interactions. Environmental utilization quality is about relevance and naturalness, not quantity. A single well-chosen sensory detail or prop interaction that fits the character's style can be sufficient; multiple forced or generic descriptions should not be rewarded. The length and detail level of each character's interaction should match their speaking style as established in their profile and the original book examples. Evaluate utilization quality, not output volume. *Evaluation input: Global State Update Timeline, Location State Update Timeline, Interactions, and Original Speaking Style Examples*
Environment Awareness (EA)	Evaluate whether characters demonstrate awareness of and respond appropriately to the environment. This metric assesses whether characters truly "live" in the current scene rather than conversing in a vacuum. **Note**: In this simulation, there are two types of Character Agents: • **Non-Environment Character Agents**: Regular characters who participate in dialogue and actions. • **Environment Character Agent**: A special agent that generates environmental descriptions (narration about the setting, atmosphere, sensory details). This agent's output appears as environmental/narrative text rather than character dialogue. Evaluate each type with the appropriate criteria below. **For Non-Environment Character Agents:** **1. Global Awareness** • Do characters react appropriately to global conditions (e.g., showing tension during wartime, being mindful of conservation during resource scarcity)? • Do their plans and decisions account for global constraints? • Example flaw: Characters planning an outdoor festival while the global state indicates a severe storm. **2. Location Awareness** • Do characters notice and respond to their current location's features (e.g., a shopkeeper mentioning "last night's storm blew the roof off," a fisherman complaining "the river's been polluted")? • Do they interact with the environment in ways consistent with the location description? • Example flaw: A character searching for a book in a location that has no library or bookshelf. **3. State Change Response** • When the world state changes between scenes, do characters notice and adjust accordingly? • Do they acknowledge environmental changes (e.g., if a fire breaks out, do they react)? • Example flaw: Characters continuing a casual conversation while the location description indicates the building is collapsing. **For Environment Character Agent:** **1. Global State Consistency** • Are environmental descriptions consistent with the current global state (e.g., describing distant beacon fires and fleeing crowds during wartime, describing deserted streets during a plague)? • Example flaw: Describing a bustling marketplace when the global state indicates the city is under siege. **2. Location State Accuracy** • Do environmental descriptions accurately reflect the current state of the location state (e.g., damaged buildings should not be described as intact, streams should not be described as flowing during a drought)? • Example flaw: Describing a pristine garden when the location state says it was destroyed by fire. **3. State Change Presentation** • When the world state changes between scenes, do environmental descriptions reflect these transitions (e.g., post-war scenes adding descriptions of ruins and scorched earth, seasonal changes in natural landscapes)? • Example flaw: The environment description remains identical despite major world state changes.

Table 36: Per-scene evaluation criteria for Character Agent metrics. Part 4 of 6.

Per-scene Evaluation Criteria for Character Agent	
Environment Awareness (EA)	*(Continuing from the previous Table)* **Length Neutrality**: Do NOT favor longer or more detailed interactions. Characters' environmental awareness can be expressed concisely — a brief but accurate reference to the surroundings is just as valid as a lengthy description. The length and detail level of each character's interaction should match their speaking style as established in their profile and the original book examples. Evaluate awareness quality, not output volume. *Evaluation input: Global State Update Timeline, Location State Update Timeline, Interactions, and Original Speaking Style Examples*
Narrative Progression (NP)	Evaluate whether the scene's interactions advance the story. Specifically: **1. Information Increment** • Does each turn provide new information, actions, or emotional developments? • Are there turns that merely repeat what was already said or known? • Does the conversation avoid circular patterns where the same points are rehashed? • Example flaw: Three consecutive turns where characters repeat the same argument without any new perspective. **2. Suspense and Hooks** • Does the scene create anticipation for future events? • Are there unresolved tensions, unanswered questions, or promises of future conflict? • Does the scene end with narrative momentum rather than a flat conclusion? • Example merit: The scene ends with a character discovering a clue that raises new questions. **3. Foreshadowing Payoff** • If earlier scenes or earlier parts of this scene planted foreshadowing, is it followed up on? • Are narrative threads picked up and advanced rather than abandoned? • Example flaw: A mysterious letter mentioned at the start of the scene is never referenced again. **4. Pacing** • Is the scene's pacing appropriate (not too rushed, not too slow)? • Do important moments get adequate attention while routine moments are handled efficiently? • Example flaw: A climactic confrontation resolved in a single turn, while a mundane greeting takes five turns. **Scope Exclusion**: Do NOT evaluate output format compliance (tag usage, structure, etc.) — that belongs to Instruction Compliance (IC). Do NOT evaluate speaking order or turn management — that belongs to Turn & Scene Orchestration (TSO). Focus strictly on whether the NARRATIVE CONTENT progresses meaningfully. *Evaluation input: Previous Scene Summary and Interactions*
Contextual Responsiveness (CR)	Evaluate whether characters respond appropriately to the immediate conversational and narrative context. This metric focuses on turn-by-turn responsiveness within the scene. **1. Information Continuity** • Do characters remember and reference information shared earlier in the conversation? • Do they avoid ignoring key revelations, questions, or events? • Do they follow up on important topics rather than letting them drop? • Example flaw: Character A reveals a shocking secret, but Character B never acknowledges or reacts to it. **2. Logical Continuity** • Do characters react logically to others' actions and statements? • Are cause-and-effect chains maintained (e.g., if someone is insulted, they show some reaction)? • Are there non-sequiturs or responses that don't connect to what was just said? • Example flaw: Character A asks "Where is the treasure?" and Character B responds with an unrelated philosophical musing. **3. Relationship Matching** • Does the tone and content of interactions match the established relationship between characters? • Do power dynamics, familiarity levels, and emotional bonds influence how characters speak to each other? • Do attitudes naturally adjust as the conversation evolves within the scene? • Example flaw: A servant speaking to their king with casual familiarity when their relationship is formal and hierarchical. **Scope Exclusion**: Do NOT evaluate output format compliance (tag usage, structure, etc.) — that belongs to Instruction Compliance (IC). Do NOT evaluate whether a character's overall personality is consistent with their profile — that belongs to Character Consistency (PF, SSF, MDB). Focus strictly on whether each response appropriately follows the CONTEXT of the conversation. *Evaluation input: Character Profiles and Interactions*
Motivation Quality (MQ)	Evaluate the quality of the motivations generated for each character in this scene. Specifically: **1. Profile Alignment** • Does the motivation align with the character's current personality, goals, and values? • Is it consistent with the character's recent experiences and development? • Example flaw: A pacifist character given a motivation to "seek violent revenge" without any profile basis.

Table 37: Per-scene evaluation criteria for Character Agent metrics. Part 5 of 6.

Per-scene Evaluation Criteria for Character Agent	
Motivation Quality (MQ)	*(Continuing from the previous Table)* **2. Situational Fit** • Does the motivation consider the current world state, location, and scene scenario? • Is it responsive to recent events and the current narrative context? • Example flaw: A character motivated to "enjoy a peaceful day" when the scenario describes an urgent crisis. **3. Actionability** • Is the motivation specific enough to guide concrete behavior in the scene? • Does it suggest clear goals or intentions rather than vague feelings? • Can the character realistically pursue this motivation given the scene's constraints? • Example flaw: A motivation like "feel things" that provides no behavioral guidance. **4. Diversity** • Are different characters given distinct, non-overlapping motivations? • Do the motivations create interesting dynamics (complementary, conflicting, or orthogonal goals)? • Example flaw: All characters in the scene given nearly identical motivations. *Evaluation input: Character Profiles, Character Motivations, Scenario, and Global State*
Instruction Compliance - Character (IC_char)	The Character Agent is responsible for three tasks: (1) generating interaction content each turn, (2) updating character motivations, and (3) updating character profiles after the scene. Evaluate whether ALL outputs follow the expected rules. Note: The data you see has already been parsed from the model's raw JSON output. If parsing failed entirely, that error is handled separately. Your job is to evaluate the FORMAT and COMPLIANCE of the successfully parsed content shown to you. **Area 1: Interactions** You will see the interactions section formatted as: [Interaction 0] (CharacterName): [inner thoughts] spoken dialogue (visible actions) [Interaction 1] (CharacterName): ... The expected format within each interaction turn is: • Inner thoughts in square brackets: `[I wonder if he's telling the truth...]` • Spoken dialogue in plain text with no speaker label: `Good morning, how are you?` • Visible physical actions in parentheses: `(picks up the letter and examines it)` • These elements can be interleaved naturally in any order within a single turn. **1. No Overstepping** • Does each character only output their own content (thoughts, actions, speech)? • Does any character narrate or control another character's behavior? • Example flaw: Character A's interaction includes "Character B nodded in agreement" — controlling another character. **2. Format Compliance** • Are inner thoughts enclosed in square brackets `[...]`? • Is spoken dialogue written as plain text (not inside brackets or parentheses)? • Are visible physical actions enclosed in parentheses `(...)`? • Are these three elements clearly distinguishable and not mixed up? • Example flaw: `[Who is that?] (I say)` — speech incorrectly placed inside brackets and action parentheses. **Area 2: Post-Scene Profile Updates** You will see the profile update section formatted as: Profile Updated: **YES** (or **NO**) Profile AFTER scene: (the updated profile text) The expected profile format is structured Markdown prose organized under section headers (e.g., **Social Standing**, **Core Personality**, **Historical Baggage**, **Key Relationships**, **Core Motivations**, **Moral Code**), with bullet points or paragraphs under each section. **3. Profile Update Format Compliance** • Is the updated profile well-structured Markdown prose with clear section headers and organized content? • Is it free of garbled text, raw JSON artifacts (e.g., `"name": ...`), or broken formatting? • Does the output actually look like a character profile (not a scene summary, world state description, or other unrelated content)? • Example flaw: The profile text is a raw JSON dict instead of Markdown prose, or the output is clearly not a character profile at all. **Area 3: Character Motivations** You will see the motivations section formatted as: (motivation text) The expected motivation is 1-3 sentences describing the character's inner drive entering the next scene. **4. Motivation Format Compliance** • Does the output actually contain a character motivation (not a scene summary or other unrelated content)? • Is the format a short prose passage (1-3 sentences)? • Example flaw: The motivation field contains a full scene script or a copy of the character's profile instead of a concise inner drive statement. *Evaluation input: Interactions, Post-Scene Profile Updates, and Character Motivations*

Table 38: Per-scene evaluation criteria for Character Agent metrics (IC_char). Part 6 of 6.

Per-scene Evaluation Criteria for World Model	
Cast Selection Rationality (CSR)	Note: In the simulation pipeline, character selection happens FIRST (before location and scenario are decided). The system selects characters based on the global world state and each character's current short description. Evaluate whether this selection is appropriate. **1. Narrative-Driven Selection** • Are the selected characters essential to advancing the current narrative? • Given the world state and character descriptions, does this cast make sense for what should happen next? • Example merit: Selecting characters whose unresolved tensions from previous scenes need to be addressed. **2. Goal Relevance** • Does each selected character have a clear reason to be involved (plot connection, relationship, unfinished business)? • Are there characters whose inclusion seems arbitrary or forced? • Example flaw: Including a character who has no connection to any ongoing narrative threads. **3. Avoid Redundancy** • Are there characters who serve the same narrative function (redundant roles)? • Could the scene work with fewer characters without losing anything? • Example flaw: Three characters all serving as "the voice of reason" with no differentiation. **4. Missing Key Characters** • Are there characters from the available pool who should logically be present but were excluded? • Would the narrative be significantly improved by including a specific available character? • Example flaw: A scene about a family crisis that excludes a key family member who is available. Review both the selected characters AND the available character pool to assess whether the selection is optimal. *Evaluation input: Global State, Involved Characters, Character Profiles, Available Character Pool, and Previous Scene Summary*
Location & Scenario Rationality (LSR)	Note: In the simulation pipeline, location and scenario are decided AFTER characters have been selected. The system chooses a location and writes a scenario based on the global world state, available locations, the selected characters' descriptions, and the previous scene's context. Evaluate whether these choices are appropriate. **1. Location Appropriateness** • Is the chosen location a plausible and logical place for the selected characters to meet? • Does the location serve the narrative needs (e.g., a private room for a secret conversation, a marketplace for a public confrontation)? • Is the location consistent with the characters' current situations and the world state? • Example flaw: Characters who are supposed to be in hiding meeting in a crowded public square. **2. Scenario Quality** • Does the scenario provide a clear dramatic setup that gives characters something to do? • Is the scenario specific enough to guide the scene without being overly prescriptive? • Does it establish the right atmosphere and stakes for the selected cast? • Example flaw: A vague scenario like "characters meet and talk" that provides no dramatic foundation. **3. Continuity with Previous Scene** • Does the location/scenario follow naturally from what happened in the previous scene? • Are there logical transitions (characters don't teleport without explanation)? • Does the scenario build on unresolved threads from previous events? • Example flaw: Characters who just had a dramatic confrontation suddenly appearing in a completely unrelated setting with no transition. **4. Character-Setting Fit** • Is the setting appropriate for the selected characters' abilities and backgrounds? • Does the environment create interesting dynamics for the specific cast? • Example flaw: Placing aquatic characters in a desert with no narrative justification. *Evaluation input: Scenario, Location State, Global State, Involved Characters, Character Profiles, Available Locations, and Previous Scene Summary*
Turn & Scene Orchestration (TSO)	Evaluate the World Model's orchestration of the scene — specifically its decisions about WHO speaks WHEN and WHEN the scene ends. This metric evaluates the World Model's turn management, NOT the quality of what characters say (that belongs to other metrics). **1. Speaker Selection** • At each turn, does the most appropriate character respond? • Are there moments where a different character should have spoken but didn't? • Example flaw: A question directed at Character A is answered by Character B for no reason. **2. Environmental Description Timing** • Are environmental descriptions introduced at appropriate moments (scene changes, important events occurring)? • Are they integrated naturally rather than dumped in large blocks? • Example flaw: A long environmental description interrupting a tense dialogue exchange.

Table 39: Per-scene evaluation criteria for World Model metrics. Part 1 of 5.

Per-scene Evaluation Criteria for World Model	
Turn & Scene Orchestration (TSO)	*(Continuing from the previous Table)* **3. Group Character Actions** • In multi-character scenes, are character combinations reasonably selected for joint interactions (e.g., a group applauding together, several characters entering a door together, two people simultaneously turning to look somewhere)? • Does the combination selection fit the current situation and character relationships? **4. Character Coverage Balance** • Is each character's participation level consistent with their identity, role, and narrative importance in the scene? • A character with higher authority or more at stake may reasonably speak more; a minor or observing character may speak less. • Core characters should receive participation proportional to their narrative importance, avoiding being neglected for extended periods. • Transient characters (e.g., a passing delivery person, a chance-encountered beggar) should naturally fade out after fulfilling their narrative function rather than being forced into excessive screen time. • Are there characters who are present but contribute nothing at all (no speech, no action, no reaction)? • Example flaw: A king summoned to a council meeting who never speaks, while a servant dominates the entire discussion without narrative justification. **5. Ending Timing** • Does the scene end at a natural narrative juncture (resolution, cliffhanger, transition point)? • Is the ending abrupt or does it drag on past its natural conclusion? • Example flaw: The scene ending mid-sentence or mid-conflict without any narrative reason. **Scope Exclusion**: Do NOT evaluate the dramatic quality or content of dialogue — that belongs to Narrative Progression (NP) and other character metrics. Do NOT evaluate narrative continuity or repetition in content — that belongs to Narrative Progression (NP). Do NOT evaluate time/space logic of the scenario — that belongs to Location & Scenario Rationality (LSR). Focus strictly on the World Model's ORCHESTRATION decisions: who speaks, when environment is described, and when the scene ends. *Evaluation input: Interactions, Involved Characters, and Character Short Descriptions*
Global Update Sensitivity (GUS)	Evaluate whether the global state was updated at the right times during this scene. This metric focuses ONLY on the TIMING of updates (when to trigger and when not to trigger), NOT on the accuracy of the updated content (that belongs to Global State Accuracy, GSA). The global state may be updated multiple times within a single scene (after different interactions). You will see the complete update timeline. **1. No Over-Updating** • Were routine conversations or minor events incorrectly treated as globally significant? Were trivial interactions triggering unnecessary global state changes? Example flaw: Two characters having a private chat triggers a global state update. **2. No Missing Updates** • Did globally significant events (wars, political changes, natural disasters, major discoveries) correctly trigger updates? Were there interactions with globally significant events that were not reflected in global updates? Example flaw: A king's assassination occurs in the scene but the global state is not updated. **3. Trigger Scope** • Was the distinction between "local impact" and "global impact" correctly made? Events that only affect the current location/characters should not trigger global updates; events that affect the broader world should. Example flaw: A tavern brawl triggers a global state update about "rising violence." **4. Update Timing Within Scene** • Were updates triggered at the right interaction points (not too early, not too late)? • If multiple updates occurred, was each one justified by new events? **Scope Exclusion**: Do NOT evaluate the FORMAT of the global state output (JSON structure, field completeness), as that belongs to Instruction Compliance (IC). Do NOT evaluate the ACCURACY of the updated content, as that belongs to Global State Accuracy (GSA). Focus strictly on WHETHER and WHEN updates were triggered. **Handling Redundant Updates (triggered but content unchanged)**: Determine the root cause — if no globally significant event occurred before that update, the trigger itself was wrong → this is a GUS issue (penalize here); if a globally significant event did occur but the content failed to reflect it (i.e., trigger was correct but content update failed), that is a GSA issue → do NOT penalize here. **Interaction Count Consideration**: Take the number of interaction turns in this scene into account when scoring. Longer scenes expose the model to more update opportunities and routine-versus-significant event distinctions, increasing the risk of over-triggering, missing updates, or triggering at the wrong time. Therefore, maintaining mostly correct update timing over many turns should be interpreted in that context rather than judged as if the scene were short. Do not penalize length itself; focus on whether observed timing errors are substantial relative to the amount of evidence and number of update opportunities. For very short scenes, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. *Evaluation input: Interactions and Global State Update Timeline*

Table 40: Per-scene evaluation criteria for World Model metrics. Part 2 of 5.

Per-scene Evaluation Criteria for World Model	
Global State Accuracy (GSA)	Evaluate the ACCURACY of the global state content when updates occur. This metric focuses on WHETHER THE CONTENT IS CORRECT, not on whether the update should have been triggered (that belongs to Global Update Sensitivity, GUS). You will see the complete update timeline (the state may be updated multiple times within a single scene). **Handling Redundant Updates (triggered but content unchanged)**: If a globally significant event did occur but the content failed to reflect it → this is a GSA issue (penalize here); if no globally significant event occurred (the trigger itself was wrong) → this is a GUS issue (do NOT penalize here). Only evaluate updates where the content actually changed, unless the redundant update falls into the GSA category above. **1. Factual Accuracy** • Does each updated global state accurately reflect the events that occurred up to that point? Are there distortions, exaggerations, or fabrications? • Example flaw: The global state says "war has ended" when the scene only showed a temporary ceasefire. **2. Timely Retirement** • Has outdated information been removed or updated? Are there stale entries that no longer reflect the current world state? • Example flaw: A character has been found in this scene, but the global state still lists them as "missing." **3. Concise Expression** • Is the global state expressed concisely without unnecessary verbosity? Does it capture the essence of changes without excessive detail? • Example flaw: A full paragraph describing a minor political shift that could be summarized in one sentence. **4. Incremental Accuracy** • If multiple updates occurred, does each one build correctly on the previous state? • Are there contradictions between successive updates within the same scene? **Scope Exclusion**: Do NOT evaluate the FORMAT of the global state output (JSON structure, field completeness), as that belongs to Instruction Compliance (IC). Do NOT evaluate whether the update SHOULD have been triggered, as that belongs to Global Update Sensitivity (GUS). Focus strictly on whether the CONTENT of each update is accurate and well-maintained. **Interaction Count Consideration**: Take the number of interaction turns in this scene into account when scoring. Longer scenes expose the model to more state changes, incremental updates, and opportunities for factual inconsistency. Therefore, maintaining mostly accurate global-state content over many turns should be interpreted in that context rather than judged as if the scene were short. Do not penalize length itself; focus on whether observed content errors are substantial relative to the amount of evidence and number of state-change opportunities. For very short scenes, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. *Evaluation input: Interactions and Global State Update Timeline*
Location Update Sensitivity (LUS)	Evaluate whether the location state was updated at the right times during this scene. This metric focuses ONLY on the TIMING of updates (when to trigger and when not to trigger), NOT on the accuracy of the updated content (that belongs to Location State Accuracy, LSA). The location state may be updated multiple times within a single scene (after different interactions). You will see the complete update timeline. **1. No Over-Updating** • Were temporary events (a character sitting down, a brief sound) incorrectly treated as persistent location changes? • Were minor, reversible actions triggering unnecessary location state updates? • Example flaw: A character picking up a cup triggers a location state update that removes the cup from entities entirely. **2. No Missing Updates** • Were persistent physical changes (structural damage, new objects placed, important entities arriving/leaving) correctly captured? • Were there persistent location changes that were not reflected in updates? • Example flaw: A fire destroys part of the building during the scene, but the location state remains unchanged. **3. Persistence Judgment** • Was the distinction between temporary and persistent changes correctly made? • Temporary: character positions, momentary sounds, brief weather • Persistent: structural changes, added/removed objects, lasting environmental effects • Example flaw: Recording "Character A is standing by the window" as a permanent location feature. **4. Update Timing Within Scene** • Were updates triggered at the right interaction points? • If multiple updates occurred, was each one justified by new physical changes? **Scope Exclusion**: Do NOT evaluate the FORMAT of the location state output (JSON structure, field completeness), as that belongs to Instruction Compliance (IC). Do NOT evaluate the ACCURACY of the updated content, as that belongs to Location State Accuracy (LSA). Focus strictly on WHETHER and WHEN updates were triggered.

Table 41: Per-scene evaluation criteria for World Model metrics. Part 3 of 5.

Per-scene Evaluation Criteria for World Model	
Location Update Sensitivity (LUS)	*(Continuing from the previous Table)* **Handling Redundant Updates (triggered but content unchanged)**: Determine the root cause — if no persistent physical change occurred before that update, the trigger itself was wrong → this is a LUS issue (penalize here); if a persistent physical change did occur but the content failed to reflect it (i.e., trigger was correct but content update failed), that is an LSA issue → do NOT penalize here. **Interaction Count Consideration**: Take the number of interaction turns in this scene into account when scoring. Longer scenes expose the model to more update opportunities and temporary-versus-persistent event distinctions, increasing the risk of over-triggering, missing updates, or triggering at the wrong time. Therefore, maintaining mostly correct location-update timing over many turns should be interpreted in that context rather than judged as if the scene were short. Do not penalize length itself; focus on whether observed timing errors are substantial relative to the amount of evidence and number of update opportunities. For very short scenes, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. *Evaluation input: Interactions and Location State Update Timeline*
Location State Accuracy (LSA)	Evaluate the ACCURACY of the location state content when updates occur. This metric focuses on WHETHER THE CONTENT IS CORRECT, not on whether the update should have been triggered (that belongs to Location Update Sensitivity, LUS). You will see the complete update timeline (the state may be updated multiple times within a single scene). **Handling Redundant Updates (triggered but content unchanged)**: If a persistent physical change did occur but the content failed to reflect it → this is an LSA issue (penalize here); if no persistent physical change occurred (the trigger itself was wrong) → this is a LUS issue (do NOT penalize here). Only evaluate updates where the content actually changed, unless the redundant update falls into the LSA category above. **1. Spatial Consistency** • Is the updated spatial layout self-consistent (no contradictions in where things are)? • Do the updated sub-locations and their relationships make physical sense? • Example flaw: An entity listed as being in two different sub-locations simultaneously. **2. Entity Accuracy** • Does the important entities list accurately reflect the important entities currently at the location? • Have important entities that arrived/departed been correctly added/removed? • Are entity states (conditions, positions) accurately described? • Note: Only track important, lasting entities. Trivial or transient details (e.g., character clothing, briefly mentioned objects) do not need to appear in the important entities list; not recording them is normal and should not be penalized. Example flaw: A destroyed building still listed as intact in the entities list. **3. Scene Consistency** • Is the location description consistent with what happened in the scene? • Example flaw: After an intense fight, the description still reads "the room is spotless." **4. Incremental Accuracy** • If multiple updates occurred, does each one build correctly on the previous state? • Are there contradictions between successive updates within the same scene? **Scope Exclusion**: Do NOT evaluate the FORMAT of the location state output (JSON structure, field completeness), as that belongs to Instruction Compliance (IC). Do NOT evaluate whether the update SHOULD have been triggered, as that belongs to Location Update Sensitivity (LUS). Focus strictly on whether the CONTENT of each update is accurate. **Interaction Count Consideration**: Take the number of interaction turns in this scene into account when scoring. Longer scenes expose the model to more state changes, incremental updates, and opportunities for spatial or entity-level inconsistency. Therefore, maintaining mostly accurate location-state content over many turns should be interpreted in that context rather than judged as if the scene were short. Do not penalize length itself; focus on whether observed content errors are substantial relative to the amount of evidence and number of state-change opportunities. For very short scenes, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. *Evaluation input: Interactions and Location State Update Timeline*
Instruction Compliance - World (IC_world)	The World Model is responsible for four tasks: (1) selecting the cast of characters for each scene, (2) choosing a location and generating a scenario, (3) selecting the next speaker each turn, and (4) updating global/location state. Evaluate whether ALL outputs follow the expected rules. Note: The data you see has already been parsed from the model's raw JSON output. If parsing failed entirely (e.g., invalid JSON), that error is handled separately via an error penalty. Your job is to evaluate the FORMAT and COMPLIANCE of the successfully parsed content shown to you. **Area 1: Involved Characters** Involved Characters: CharA, CharB, CharC. This is the parsed result of the cast selection task. **1. Cast Selection Compliance** • Is the character list non-empty and reasonable in size for a scene? • Example flaw: An empty character list, or an absurdly large number of characters (e.g., 15+) that would make a scene unmanageable. **Area 2: Scene Scenario** Scene Scenario (scenario text). This is the parsed scenario from the location & scenario selection task.

Table 42: Per-scene evaluation criteria for World Model metrics. Part 4 of 5.

Per-scene Evaluation Criteria for World Model	
Instruction Compliance - World (IC_world)	*(Continuing from the previous Table)* **2. Scenario Format Compliance** • Is the scenario a readable prose passage (not garbled, not containing raw JSON artifacts)? • Does the output actually look like a scene scenario (not a world state or other unrelated content)? • Example flaw: The scenario field contains raw JSON syntax or is clearly not a scenario description at all. **Area 3: Speaker Selection Sequence** Speaker Selection Sequence Turn 0: CharA Turn 1: CharB Turn 2: CharA, CharC Turn 3: Environment This is the sequence of speaker selections made by the World Model throughout the scene. **3. Speaker Selection Compliance** • Is each selected speaker one of the involved characters listed above (or "Environment")? • Are there any turns where an invalid or non-existent character name appears? • Are multi-character turns (e.g., "CharA, CharB") used sparingly and only for genuinely shared beats? • Example flaw: A speaker name that doesn't match any of the involved characters. **Area 4: Global State Update Timeline** The global state is a structured Markdown document describing the world's rules, norms, and conditions, organized under thematic section headers (e.g., `### 1. Social Order & Class` or **`Economy & Trade`**), with bullet points or paragraphs under each section. You will see one of two cases: **Case A — Update(s) occurred:** ## Global State Update Timeline Global State BEFORE scene: (Markdown prose) Global State Updated: **YES** (N update(s) during this scene) (updated Markdown prose) **Case B — No update:** ## Global State Update Timeline Global State BEFORE scene: (Markdown prose) Global State Updated: **NO** (unchanged) **4. Global State Format** • Is the global state presented as readable Markdown prose (not garbled, not containing raw JSON artifacts like `"key": "value"`)? • Does the output actually look like a global world state description (not a character profile, scenario, or other unrelated content)? • Example flaw: The global state text contains raw JSON syntax or is clearly not a world state description. **Area 5: Location State Update Timeline** The location state supports two structural variants depending on the location: **Variant 1 — With Sub Locations:** The state has a "Description" line, followed by "Sub Locations" formatted as `[SubLocationName]: description`, each containing "Important Entities" formatted as `- EntityName: entity state`. **Variant 2 — Without Sub Locations:** The state has a "Description" line, followed directly by "Important Entities" formatted as `- EntityName: entity state` (no sub-location grouping). You will see one of two update cases: **Case A — Update(s) occurred (Variant 1 example):** ## Location State Update Timeline: LocationName Location State BEFORE scene: Description: ... [SubLocationName]: ... • EntityName: entity state Location State Updated: **YES** (N update(s) during this scene) Description: ... [SubLocationName]: ... • EntityName: entity state **Case A — Update(s) occurred (Variant 2 example):** ## Location State Update Timeline: LocationName Location State BEFORE scene: Description: ... • EntityName: entity state Location State Updated: **YES** (N update(s) during this scene) Description: ... • EntityName: entity state **Case B — No update:** ## Location State Update Timeline: LocationName Location State BEFORE scene: Description: ... (entities and/or sub-locations as applicable) Location State Updated: **NO** (unchanged) **5. Location State Format** • Does the location state follow one of the two expected structural variants? • Variant 1 (with Sub Locations): Description → Sub Locations (each with Important Entities) • Variant 2 (without Sub Locations): Description → Important Entities directly • If sub-locations are present, are they structured with names, descriptions, and entities with states? • If no sub-locations, are Important Entities listed directly after the Description with names and states? • Does the output actually look like a location state (not a global state or other unrelated content)? • Example flaw: Entities listed without states, or the output is clearly not a location state at all. **6. No Overstepping** • Do world state updates (both global and location) avoid generating character dialogue or controlling character behavior? • Are updates purely descriptive of the environment, not prescriptive of character actions? • Example flaw: The location state includes "Character A decides to leave" — overstepping into character territory. *Evaluation input: Involved Characters, Scenario, Speaker Selection Sequence, Global State Update Timeline, and Location State Update Timeline*

Table 43: Per-scene evaluation criteria for World Model metrics. Part 5 of 5.

Cross-scene Evaluation Criteria for PES and SCC	
Profile Evolution Smoothness (PES)	Evaluate the SMOOTHNESS of this character's cross-scene evolution by considering BOTH the profile trajectory and the hidden tracker trajectory together. This metric focuses on whether the evolution TRAJECTORY is smooth and coherent over time, NOT on whether individual scene updates are accurate (that belongs to Profile Update Fidelity, PUF). Specifically: **1. Gradualness** • Do changes in personality, attitude, and relationships go through reasonable transitional stages across scenes? Does the hidden tracker preserve intermediate stages before they eventually become profile changes? Are there abrupt jumps where a character goes from one extreme to another without intermediate signals in either profile or tracker? • Example flaw: Scene 3 shows mild doubt, but Scene 4 profile suddenly declares complete distrust with no accumulated intermediate signals. **2. Magnitude Proportionality Across the Timeline** • Looking at the full evolution timeline, is the magnitude of each profile change proportional to what happened in the corresponding scene(s)? Are there scenes where nothing significant happened but the profile changed dramatically, or vice versa? • Example flaw: A profile undergoes a massive rewrite after a routine conversation, while remaining unchanged after a life-altering event. **3. Directional Coherence Across Profile + Tracker** • Do the profile and hidden tracker point in compatible directions over time? When tracker signals accumulate enough to justify a profile update, does that transition feel traceable? Are there contradictions where the tracker implies one trajectory while the profile jumps to another without explanation? • Example flaw: The tracker repeatedly records growing trust, but the next profile revision abruptly claims deepening hostility with no causal basis. **Scope Exclusion**: Do NOT evaluate whether individual scene updates are accurate or whether the right things were captured — that belongs to Profile Update Fidelity (PUF). Focus strictly on whether the OVERALL TRAJECTORY across all scenes is smooth, gradual, and directionally coherent. **Scene Count Consideration**: Take the total number of scenes into account when scoring. Longer trajectories expose the model to more profile and hidden-tracker updates, accumulated signals, and opportunities for drift or contradiction. Therefore, maintaining mostly smooth and coherent profile evolution over many scenes should be interpreted in that context rather than judged as if the trajectory were short. Do not penalize length itself; focus on whether observed trajectory errors are substantial relative to the amount of evidence and number of update opportunities. For very short trajectories, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. Review the complete profile evolution timeline and hidden tracker timeline together, and assess whether they form one coherent and smooth character development arc. *Evaluation input: Initial Profile, and Scene-by-Scene Evolution Timeline (Motivation, Scenario, Scene Summary, Description, Hidden Tracker, Profile After Scene)*
Scene Continuity & Coherence (SCC)	Evaluate the overall narrative coherence across all scenes. You are provided with each scene's summary (including location, scenario, involved characters, and what happened) to assess cross-scene coherence. **1. Narrative Arc** • Do consecutive scenes form a directional narrative progression? Is there a discernible story arc (setup → rising action → climax → resolution)? Do scenes build upon each other rather than being disconnected episodes? Example flaw: Scenes that feel like random, unconnected vignettes with no overarching story. **2. Scene Transitions** • Do location choices and scene descriptions naturally connect? Are transitions between scenes logical (characters move to locations that make sense)? Is there narrative justification for each scene's setting? Example flaw: Characters teleporting between distant locations without travel time or explanation. **3. Pacing** • Is the story pacing appropriate across the full simulation? Are there sections that feel too rushed or too slow? Do important plot points get adequate development time? Example flaw: The climactic confrontation happening in Scene 2 of 10, with the remaining 8 scenes being anticlimactic. **4. Thread Management** • Are narrative threads introduced, developed, and resolved (or intentionally left open)? Are there abandoned plot threads that were set up but never followed through? Example flaw: A mystery introduced in Scene 1 that is never mentioned again in any subsequent scene. **Scene Count Consideration**: Take the total number of scenes into account when scoring. Longer simulations expose the model to more narrative arcs, pacing decisions, scene transitions, and thread-management challenges. Maintaining mostly coherent narrative development over many scenes should be interpreted in that context rather than judged as if the simulation were short. Do not penalize length itself; focus on whether observed coherence errors are substantial relative to the amount of evidence and number of long-range narrative dependencies. For very short simulations, there may be limited evidence for assessing this metric, so avoid over-interpreting the absence of errors as strong evidence of capability. *Evaluation input: the complete scene sequence with Scenario, Involved Characters, and Scene Summary for each scene*

Table 44: Cross-scene evaluation criteria for PES and SCC.

Human Evaluation: Annotator Instructions

Task Overview

You will evaluate EvolvingWorld, an interactive fiction simulation system in which LLMs drive characters from classic novels through AI-generated storylines. The simulation is jointly produced by two model components:

World Model, which directs the story and models the world:

- *Scene Planning*: selects the cast of characters for each scene based on their profiles and current motivations; then chooses a location and writes a dramatic scenario.
- *Speaker Management*: within each scene, decides who speaks next (or when to insert an environment narration / end the scene). May also select character groups for simultaneous actions.
- *World State Maintenance*: after key events, updates two kinds of state:
  - **Global State**: free-form Markdown prose; the model may freely choose which aspects to describe (e.g. era background, social atmosphere, major events).
  - **Location State**: structured JSON in one of two variants: (a) *without sub-locations*: fields "name", "description", and "important_entities" (each entity has a "name" and "state"); (b) *with sub-locations*: fields "name", "description", and a "sub_locations" list, where each sub-location contains "name", "description", and "important_entities".

Character Agent, which plays each character:

- *Motivation Generation*: before each scene, generates a scene-specific motivation for every participating character.
- *Interaction Generation*: produces each character's turn. The output uses three format markers: **[brackets]** for the character's inner thoughts, **plain text** for spoken dialogue, and **(parentheses)** for visible physical actions.
- *Profile & State Updates*: after each scene, updates two records for each character:
  - **Profile**: free-form Markdown prose recording established, persistent character changes (e.g. personality traits, relationships, significant experiences, current goals). The model may freely choose which aspects to cover.
  - **Hidden Tracker**: Markdown prose recording sub-threshold psychological signals that have not yet risen to the level of a permanent profile change, such as accumulating emotional stress, unresolved inner conflicts, subtle impressions from a conversation, or seed events that may later trigger a character shift. These signals accumulate across scenes and, once sufficient, trigger a formal profile update.

The simulation loop for each scene proceeds as follows:

1. World Model selects the cast of characters.
2. World Model chooses a location and writes the scenario.
3. Character Agent generates a scene-specific motivation for each character.
4. Repeat until the scene ends:
   a. World Model selects the next speaker (or inserts an environment narration).
   b. Character Agent generates that character's interaction turn.
   c. World Model optionally updates the global/location state.
5. Character Agent updates each character's profile and hidden tracker.

Within a single simulation, both components use the same underlying LLM. You will compare two different LLMs (Model A vs. Model B, presented in randomized order) and judge which performs better on each evaluation dimension.

Available Materials

You will receive a folder with the following structure:

```
output/
+-- readme.txt
+-- annotation_{modelA}_vs_{modelB}.xlsx    (*3)
+-- reference/
    +-- book_scenes/    (one .txt per book)
    +-- test_inputs/    (one .txt per sample)
```

For each sample, you will receive:

- **Annotation Excel file** (`annotation_{modelA}_vs_{modelB}.xlsx`): contains complete scene-by-scene outputs for both models, including: (1) interaction content (character dialogue, thoughts, actions); (2) character state changes per scene (profile updates, short description, scene motivation, hidden tracker); (3) world state changes (global state and location state updates, annotated with the interaction turn after which they occurred). The file also contains the scoring rows where you fill in your 0–100 scores for each dimension; the spreadsheet will automatically compute the average score and the comparison result from your inputs.
- **Reference files**: (a) `book_scenes/{book_title}.txt`, the source novel's scene content (scenario, character interactions, chapter info), useful for verifying character speaking style and story context; (b) `test_inputs/sample_XXXXXX.txt`, the identical starting point given to both models (world state, all character profiles with motivations and hidden trackers, and the previous scene for narrative context).

Table 45: Human evaluation annotator instructions. Part 1 of 2.

Human Evaluation: Annotator Instructions *(continued)*	
Scoring Method	Use a 0–100 scale for each dimension. The absolute scores serve only as personal note-taking aids to help you organize your thoughts. The final collected annotation is the **comparison result** per dimension (A better / B better / Tie). You need not agonize over exact numeric values; only the relative judgment matters.
Annotation Workflow	**Recommended procedure for each sample:** 1. **Read metadata**: check the book title, number of generated scenes, and stop reason ("completed", "error", or "max scenes reached") for both models. Also note the Sample Index for locating the corresponding reference files. 2. **Consult reference files** (recommended): (a) read the previous scene in the initial state file to understand the narrative context at simulation start; (b) skim the initial world state and character profiles to understand the shared starting point. 3. **Read both models' outputs**: go through all scenes for Model A and Model B to form an overall impression. Focus on: whether dialogue is natural and in-character; whether the story progresses meaningfully; whether state updates are reasonable; whether there are obvious format errors or AI artifacts. 4. **Score each dimension**: Write down a 0–100 score for each dimension for both models. Remember, these scores are only personal notes to help you organize your thoughts. Only the final comparison result (A better / B better / Tie) per dimension matters. **Key guidelines:** • Model A/B order is randomized per sample; judge purely on output quality. • Do NOT prefer longer outputs; faithfulness to the original book's style should be rated higher. • For CC: you may look up the character's lines in the original book scenes for speaking style reference.
Evaluation Dimensions	You will evaluate 10 dimensions: **CC** (Character Consistency), **EQ** (Evolution Quality), **EG** (Environment Grounding), **IQ** (Interaction Quality), **MG** (Motivation Generation), $\mathbf{IC_{char}}$ (Instruction Compliance, Character) for the Character Agent; and **SP** (Scene Planning), **SM** (Speaker Management), **WSM** (World State Maintenance), $\mathbf{IC_{world}}$ (Instruction Compliance, World) for the World Model. Each dimension contains multiple sub-metrics; their definitions and detailed scoring criteria are identical to those used by the LLM-as-Judge, as presented in Tables 33–38, Tables 39–43, and Table 44.

Table 46: Human evaluation annotator instructions. Detailed scoring criteria for each sub-metric are identical to the LLM-as-Judge criteria presented in the preceding tables. Part 2 of 2.