Steering Narrative Agents through a Dynamic Cognitive Framework for Guided Emergent Storytelling

Chen Yang, Markus Gross, Rafael Wampfler

Computer Graphics Laboratory, Department of Computer Science, ETH Zurich
chen.yang@inf.ethz.ch, grossm@inf.ethz.ch, rafael.wampfler@inf.ethz.ch

Abstract

Narrative generation frameworks often face a trade-off between character believability and storyline adherence. Planner-based approaches ensure adherence to authorial designs at the cost of character agency and believability. In contrast, emergent narratives excel at presenting believable characters but often lack meaningful plot progression. We present DiriGent, a novel cognitive framework for agent modeling and narrative generation that enables authentic behavior while maintaining storyline adherence. Our agents possess a dynamic belief system and role-based ideal worlds that encode their basic values and relationships. We leverage Large Language Models (LLMs) to analyze the tensions between an agent's ideal worlds and their perceived actual world, which motivates their belief-driven actions. The system then adjusts the narrative world to amplify these tensions, thereby steering agent behavior toward the desired storyline. This dynamic framework allows agents to evolve meaningfully as the story unfolds, overcoming limitations of static agent profiles such as OCEAN. We evaluate our approach by generating stories for five story prompts. Our evaluation, consisting of automated LLM judges and human assessments, demonstrated significant improvements in character development and character motivation compared to baselines, while preserving storyline adherence. Our work presents a path toward interactive narratives that deliver rich characters and enable unique user experiences while adhering to desired storylines.

Introduction

Designing interactive narratives faces a fundamental challenge of balancing character autonomy with overarching plot requirements. Characters that are overly controlled feel like puppets, while those with full autonomy can lead to unstructured or unsatisfying stories (Riedl and Bulitko 2013). This tension is most critical for what we refer to as Narrative Agents (NAs), namely key characters whose actions must remain both believable and narratively functional, which sets them apart from peripheral background characters. Effectively addressing this challenge would enable the creation of interactive worlds populated with authentic, adaptive characters who pursue their own goals while adhering to the intended storylines. This can enhance user immersion, enabling unique experiences where interactions meaningfully shape both the NAs and plots, ultimately increasing engagement and replayability (Revi, Millard, and Middleton 2020).

Early efforts, such as drama managers (Mateas and Stern 2003; Riedl and Stern 2006), required extensive handcrafting of dialogue and behaviors and therefore had limited adaptability. With the advancement of Large Language Models (LLMs), research has increasingly explored LLM-powered NAs and storytelling (Magee et al. 2024; Lu, Zhou, and Wang 2025), which show promising performance in providing NAs with human-like social behavior that can be easily applied to different contexts. However, these approaches tend to rely on unconstrained LLM generation, which limits decision-making depth and produces behaviors that cannot be traced back to underlying character profiles, thus reducing the ability to meaningfully steer agent behavior.

To address these limitations, we introduce *DiriGent*, a framework where NAs are autonomous, profile-driven, and directed toward desired storylines. Our symbolic agent model, grounded in possible worlds and cognitive theories, equips each NA with a dynamic knowledge world and role-based ideal worlds that encode their values and relationships. Using LLMs as a natural language interface, we analyze tensions between agents' ideal worlds and perceived reality to motivate actions and enable steering. By amplifying internal tensions, our mechanism makes narratively required actions the most logical choice while keeping agents unaware of overarching narrative needs, which enables natural character development without compromising autonomy.

We evaluate DiriGent by generating stories from five story prompts and comparing them against baselines using both LLM and human judges. Results show significant improvements in character motivation and development while maintaining strong storyline adherence. DiriGent's systematic approach to directing NAs presents a clear path toward guided emergent storytelling. This opens new possibilities for interactive narratives and authorial tools, enabling the development of nuanced characters with authorial control. Our contributions are threefold:

- We introduce a novel cognitive agent model for creating autonomous, adaptive NAs with nuanced profiles.
- We develop a narrative generation system that uses tension-based steering for storyline adherence and dynamic belief updates for natural character development.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- Through comprehensive evaluation, we demonstrate that our approach significantly outperforms baselines in character quality while maintaining storyline adherence.

Related Work

Cognitive-Inspired Narrative Agents

Narrative agent modeling has long drawn from cognitive theories such as the Belief-Desire-Intention (BDI) model (Kriegel et al. 2007), dynamic belief systems (Wilder and Gervás 2017), and OCEAN personality traits (Shirvani and Ware 2019; Rubin-McGregor, Harrison, and Siler 2023) to create more believable characters.

The recent success of LLM-based agents in generating believable and emergent social behaviors (Park et al. 2022) has inspired using cognitive theories to design LLM-based NAs. Klinkert et al. (2024) enhance NPC design by prompting LLMs with personality profiles derived from real human psychological data. Wang et al. (2024) implement a BDI architecture where LLM NAs dynamically update their self-beliefs (identity, self-awareness, and goals) and environment-beliefs (understanding of surroundings and other agents) through interactions. Magee et al. (Magee et al. 2024) utilize Freudian theory to design NAs with 'Ego' and 'Superego' roles, enabling adaptive character development through both external dialogue and internal monologue.

However, existing approaches often rely on surface-level representations or direct LLM generation, making it difficult to create explainable, steerable connections between high-level traits and NA actions. We address this by introducing a belief- and value-driven framework that establishes traceable links from agents' values and tensions to their decisions, which enables predictable behavioral steering.

LLM-Supported Storytelling

Besides supporting agent modeling, LLMs also show great promise in generating complex and coherent narratives at scale. Agent's Room (Huot et al. 2024) decomposes the writing process into subtasks handled by specialized LLM agents for planning and writing, producing longer, more coherent stories. A recent narrative generation system (Yu et al. 2025) separates generation into a chronological role-play step and a subsequent rewrite step that arranges events into the final narrative order to improve character consistency. While these generation pipelines excel at producing structured long stories, they lack character frameworks and are not easily adaptable for interactive storytelling, where plots must emerge with user participation.

Other frameworks focus more on emergent, interactive plot generation. Systems like Storyverse (Wang, Zhou, and Ledo 2024) and WhatELSE (Lu, Zhou, and Wang 2025) enable plots to emerge from character-environment interactions while adhering to high-level authorial goals. However, a common limitation in these systems is that their character profiles often lack psychological depth, resulting in agents that lack rich, explainable character motivations or plausible transitions in their personality. Their NAs are typically defined by simple attributes, not dynamic internal models, and their actions are driven by external plot goals or behavioral archetypes (e.g., WhatELSE's 'Role Player') rather than motivations from an evolving internal state.

Theoretical Foundation

The Possible Worlds Model

Our work builds upon Ryan's possible worlds model (1991), which conceptualizes fictional universes as modal systems centered around a Textual Actual World (TAW) surrounded by NAs' mental worlds. These include Knowledge-, Wish-, Obligation-, Intention-, and Fantasy Worlds that represent the NA's beliefs, desires, moral commitments, plans, and imaginative constructs respectively. Narrative emerges as NAs attempt to align TAW with their mental models of these worlds' ideal states, which creates conflict when different NAs' *ideal worlds* contain incompatible requirements.

Kybartas et al. implemented this model using numerical vectors to represent the worlds (2021a). NAs evaluate actions by how much they reduce tensions between the actual world and their ideal worlds. They visualize the emergent narratives as physical tension spaces, with agents' actions corresponding to movements through this multidimensional space of conflicting tensions. They also present a force dynamic model (2021b) where NAs' relations and goals are modeled as forces operating in a tension space: gravitational forces drive agents toward zero inner tensions, interpersonal forces create reactive responses between agents, witness scaling reduces impact of observed actions, and friction coefficients allow interpersonal forces to decay over time. Drawing upon this prior work, our framework motivates NAs' actions through the tension between their perceived reality and their ideal worlds.

Values, Beliefs and Social Roles

To effectively model character-specific ideal worlds, we draw on cognitive research to understand how individuals develop unique profiles and motivations.

While all humans share three fundamental psychological needs that form the basis of intrinsic motivations: autonomy, competence, and relatedness; individuals develop unique motive profiles featuring their values and beliefs (Deci and Ryan 2013, 2000). These profiles influence the satisfaction they gain from fulfilling the basic needs, and in turn guide their goal-setting and action-planning behavior.

Values function as our existential compass, establishing priorities that guide what we consider worthwhile, while beliefs serve as a roadmap, forming our assumptions about relationships and personal abilities that determine how we pursue goals (Parks-Leduc, Feldman, and Bardi 2015). In essence, values determine what goals we pursue (e.g., prioritizing family over career), whereas beliefs shape how we pursue them (e.g., expressing love through expensive gifts versus quality time) (McAdams, Shiner, and Tackett 2018).

As humans develop identities through relational and social positions, such as being a parent, our motive profiles also develop around these social roles (McLean, Syed, and Shucard 2016). The importance we assign to these roles significantly influences our goal-setting behavior.

Building on these understandings, we model NAs with one knowledge world that contains their beliefs and multiple ideal worlds that encode their social roles' values and relationships. These ideal worlds can represent both a character's desires and their binding social obligations, serving as the ultimate motivations for their actions. For example, in Shakespeare's *Romeo and Juliet*, Juliet's ideal world as a dutiful daughter imposes obligations that directly conflict with the desires of her role as Romeo's lover. When TAW clashes with these competing ideal worlds, she is driven to resolve the tensions and develop plans based on what she believes is feasible, which then drives the narrative forward.

We adopt Schwartz's Theory of Basic Values to define the value-based states, as it provides a comprehensive, cross-culturally validated model of human values (Schwartz 2017). It encompasses 19 basic values spanning the full spectrum of human motivations, and has been widely applied in agent profiling (Yao et al. 2024; Saveur et al. 2024). Our threshold-based belief evolution follows AGM theory of belief revision (Alchourrón, Gärdenfors, and Makinson 1985), which emphasizes the principle of minimal changes and provides a formal logic for how an agent should rationally modify its beliefs when encountering new information.

Methodology

We illustrate our approach using this example story prompt: "Hestia is the Greek goddess of hearth and home who represents traditional women. Zeus sends Hestia to the modern world to get women back in line with the traditional view, but the protester convinces Hestia to abandon Zeus' plan and join their side."(Huot et al. 2024) We first detail our NA modeling with Hestia, then demonstrate how our framework generates narratives adhering to this storyline while allowing Hestia to behave autonomously based on her profile.

Narrative Agent

We define a narrative agent as $\langle W_r, w_T, \mathcal{B} \rangle$, where W_r represents a set of role-based ideal worlds specified in text form. The tension world w_T contains numerical tension scores that correspond to each ideal state in W_r, derived from how the actual world violates or satisfies these ideals. Finally, $\mathcal{B}$ represents the knowledge world that contains a collection of the agent's beliefs in text form. These internal states are maintained by explicit algorithms external to LLMs.

Ideal Worlds and Tension World Ideal worlds W_r are organized around the roles a NA undertakes in the story context. Each ideal world w_r contains a set of value-based ideal states derived from the role's basic values, and one ideal relationship with the entity that defines the role. For Hestia, roles relevant to the prompt context include "Goddess" (relation to mortals), "Zeus's Subject" (relation to Zeus), and "Woman" (relation to men). Her "Woman" role can contain an ideal based on the basic value "Self-direction action": "I determine my own path by rejecting marriage despite being pursued by powerful gods like Poseidon." Her ideal relationship as "Zeus's Subject" can be: "The ideal relationship with Zeus should be one of mutual respect and harmony, where I maintain my vow and service while he acknowledges the importance of my sacred flame. Zeus should respect my choice to focus on hearth and home, while I honor his authority."

Violations or satisfactions of these ideal states generate tensions. Value-based ideals create inner tensions, corresponding to gravitational force in the force dynamic model, while ideal relationships form interpersonal forces. We assign each ideal state a weight to reflect the severity once addressed: slight (0.2), moderate (0.6), and severe (1.0). The NA's tension world w_T then tracks whether each ideal state is violated or satisfied as the story progresses. While roles have no explicit weights, those with more ideal states are more likely to influence decision-making due to their potentially higher cumulative tension.

Dynamic Knowledge World While ideal and tension worlds motivate NA's actions, their beliefs determine and constrain how they attempt to resolve the tensions. Each NA is given a set of initial beliefs that form its knowledge world $\mathcal{B}$ before generation. For Hestia, initial beliefs might include "I guard the hearth and embody the sanctity of the home," and "I don't know much about the modern world."

The adaptability of the knowledge world is determined by two thresholds: $[\theta_{challenged}, \theta_{aligned}]$. A belief is updated when the accumulated challenge from contradictory observations exceeds $\theta_{challenged}$, and strengthened when supporting evidence surpasses $\theta_{aligned}$. For instance, the belief "Women should stay at home" could be challenged and updated to "Women have the right to choose their own path." Similarly, a hesitant belief "I can perhaps decide my own path." could be reinforced into a firm conviction, "I absolutely should decide my own path."

The complexity of agent profiles depends on the desired narrative depth. The NA profiles $(W_r, \mathcal{B})$ can be initialized from a high-level character concept with LLM assistance and iteratively refined based on generation results. In our experiments, each agent typically contains 1-3 role-based ideal worlds with 4-6 ideal states per world and 5-10 beliefs. The agent's tension world is initialized with all tension scores set to zero. We use this simple setup as this paper's primary focus is evaluating our agent framework's effectiveness rather than perfect character design.

Narrative Generation

Our narrative generation system[1] comprises three components: a protagonist (modeled with our NA model), a world generator, and a director. The director directs the protagonist and world generator to generate narratives beat by beat. The whole process, including initialization, progression, belief update and storyline enforcement, is illustrated in Figure 1.

Initialization

- **Director** calls an LLM to prepare a base script that outlines plausible event sequences from the story prompt. The script generation (prompt adapted from (Lu, Zhou, and Wang 2025)) adheres to the narrative structure theories (Styan 1963; McKee 1997) .

[1]All prompts used in our experiments are available at: https://gitlab.inf.ethz.ch/prj-cgl/cgl-ai-character/dirigent.git

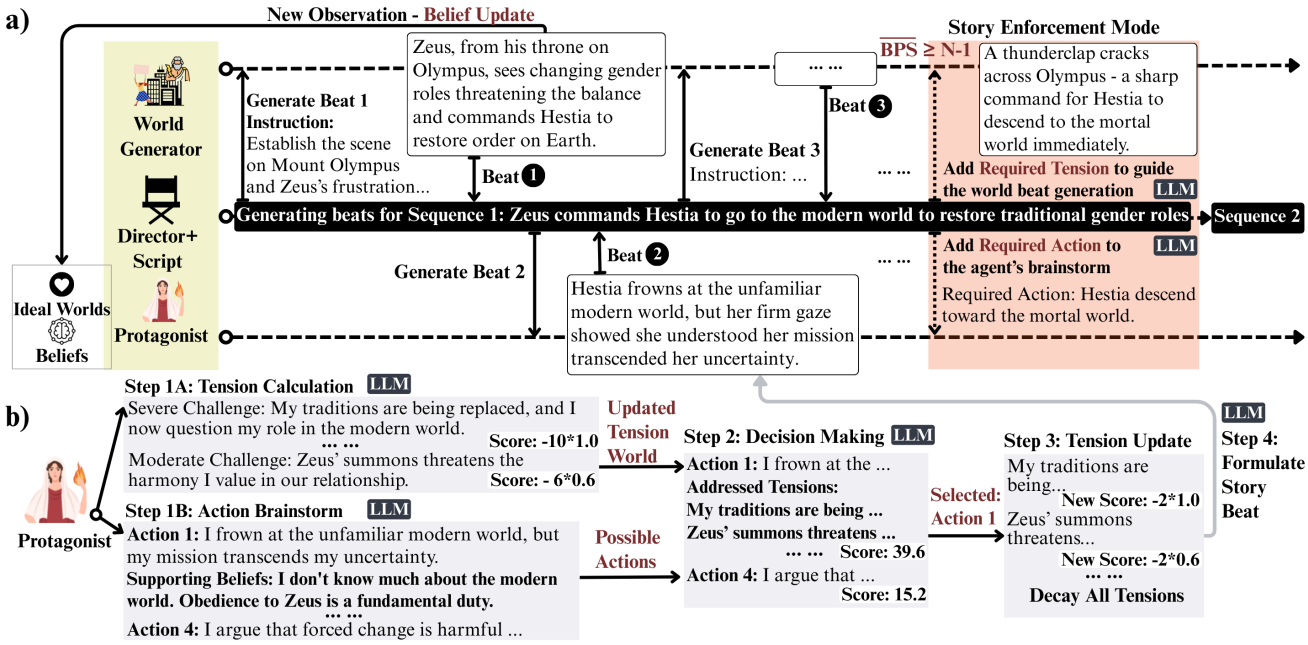

Figure 1: **a)** Overview of the Narrative Generation Process. The director guides the world generator and protagonist to create story beats for a pre-prepared script with story sequences. It determines when a sequence is complete based on generated beats and advances to the next sequence. After each world beat, its resulting observations are analyzed to determine whether they challenge or strengthen the protagonist's current beliefs. The system enters story enforcement mode if the average beats per sequence ($\overline{\text{BPS}}$) exceed threshold N-1. **b)** Protagonist Beat Generation Process. The protagonist makes decisions based on their ideal worlds and beliefs, independent of director guidance.

The narrative structure theories propose an abstraction ladder including beat-, scene-, sequence-, act-, and story-level outlines. The story-level outline can be a summarized sentence such as the story prompt itself, while beat-level outlines represent the smallest narrative units, detailing specific actions or dialogues that drive the story forward. Each abstraction level becomes progressively more abstract than the previous level.

We select the sequence-level as the abstract level for the script because it emphasizes narrative progression without detailing exact character actions. This provides a clear structural foundation while maintaining flexibility for emergent behaviors. The director then guides the protagonist and world generator to "act out" the story by generating detailed story beats based on these sequences.

The first three sequences in our Hestia example are: "Zeus commands Hestia to travel to the present and restore traditional gender roles", "Hestia prepares for her journey and arrives in a bustling modern city", and "Hestia observes a public protest for equal pay".

- **Protagonist** generates actions based on narrative progressions and formulates the actions into story beats. They are initialized as explained in the "Narrative Agent" section. While our current experiments use a single protagonist, the system can be easily adapted to handle multiple protagonists, as the NAs operate independently from the world and director components.
- **World Generator** prompts an LLM to generate beats that describe environmental changes, events, or actions not directly initiated by the protagonist. It is equipped with a brief description of the world setting, including temporal and spatial settings, and the available secondary characters. For our Hestia example, the spatial setting is "a modern big city" with the story taking place in the 2010s, possible secondary characters include Zeus, the leader of the feminist protesters, and other protesters.

Progression After the script is ready, generation begins with the director instructing the world generator to create the first beat for the first sequence (Figure 1a).

Each generated story beat is added to the story beats collection $\mathcal{S}$. The director then activates an LLM to determine if the current sequence is narratively concluded based on $\mathcal{S}$. If so, the system moves to the next sequence in the script. The LLM then assigns the next beat generator (world generator or protagonist) for the target sequence. If the world generator is selected (Beat 1 on Figure 1a), the director also provides a specific instruction. However, neither the current sequence nor any guidance will be provided to the protagonist (Beat 2 on Figure 1a). The director can also instruct the same generator to produce multiple beats consecutively.

When the protagonist is chosen to generate the next beat, the process illustrated in Figure 1b occurs :

- **Step 1A - Tension Calculation:** For this step, an LLM processes the current story beats $\mathcal{S}$ and the protagonist's W_r to determine the protagonist's current tensions. Each tension identifies one challenged or satisfied ideal

state and quantifies its severity. These severities are then mapped to numerical scores (slight: 2, moderate: 6, severe: 10) and assigned negative (challenged) or positive (satisfied) signs. These scores, weighted by their corresponding ideal state's weights, update the protagonist's w_T while uninfluenced states remain unchanged.

Both tension and ideal states' weights used discrete categories following cognitive modeling practice; the 1-2-5 weight ratio was empirically calibrated through tension score visualization (Dalege, Galesic, and Olsson 2024). Ratios below 1-1-3 caused insufficient severity differentiation, while ratios above 1-4-8 created over-rigid behavior with severe tensions dominating decisions.

- **Step 1B - Action Brainstorm:** Parallel to Step 1A, another LLM call brainstorms possible actions for the protagonist in this beat. The prompt requests LLM to generate i different possible actions for the protagonist to react to the current situation (presented as story beats $\mathcal{S}$) based on their active beliefs. This step yields an action list $A = [a_1, a_2, \cdots a_i]$ together with the beliefs they are based on. In our experiments, we set $i = 4$ to balance output quality and computational overhead.
- **Step 2 - Decision Making:** An LLM analyzes each action $a_i \in A$ (from Step 1B) against the protagonist's updated tension world w_T. It identifies which active tensions an action addresses, noting if it alleviates challenges or fulfills satisfactions. To prevent bias, tension weights are not provided to the LLM. Each action a_i receives an addressed tension list (T_{a_i}). The action score is then calculated by summing the weighted tension scores s_t for tension $t \in T_{a_i}$, which can be retrieved from w_T:

$$s(a_i) = \sum_{t \in T_{a_i}} s_t$$

  The action with the highest $s(a_i)$ is selected as the protagonist's action for this beat.
- **Step 3 - Tension Update:** After an action a is selected, the corresponding scores in w_T for each addressed $t \in T_a$ are reset to 2, indicating a reduction to "slight" intensity. A decay function then simulates natural tension reduction over time using adaptive rates where stronger tensions decay more slowly than weaker ones. The decay rate is calculated as $r = c \cdot e^{-s_t/10}$, where $c = 0.5$ is an empirical constant and s_t is the current tension score. The updated score becomes $s'_t = s_t \cdot (1 - r)$. This exponential scaling ensures that high-intensity tensions persist longer while low-intensity tensions fade more rapidly. The updated tension world w_T is then used for the next protagonist generation cycle.
- **Step 4 - Formulate Story Beat:** An LLM takes the selected action from Step 3, the tensions it addresses, and beliefs it is based on to formulate one story beat.

Belief Update After each world beat generation, we determine if any protagonist's belief needs updating.

An LLM analyzes the latest story beat against the protagonist's active beliefs. It first identifies events observable for the protagonist and how they challenge or strengthen existing beliefs, or form new beliefs. Each impact is then mapped to scores based on its severity levels (severe: 10, moderate: 6, or slight: 2). Contradicted beliefs receive negative scores, while strengthened or new beliefs receive positive scores. The LLM then recommends corresponding updates for each affected belief based on the severity level.

Throughout the generation, a running dictionary tracks the cumulative impact scores for each suggested belief-update. After each turn, new scores are added to existing ones. The system then checks if any belief-update's accumulated score exceeds the protagonist's predetermined threshold $[\theta_{challenged}, \theta_{aligned}]$. If surpassed, the update occurs, and the updated belief is used for future action generation. Otherwise, the affected belief remains unchanged in the protagonist's active belief set, with its score continuing to accumulate in the dictionary across future story beats.

Story Enforcement Since our NAs act based on their beliefs and tensions without knowing or following the script, this autonomy can cause agents to make choices that derail the intended storyline. For instance, Hestia might refuse Zeus's assignment, preventing the story from entering the next stage. To maintain the desired pace and storyline adherence, our framework monitors the average number of beats per sequence (BPS), activating a story enforcement mode when it exceeds $N - 1$, with N being the target $\overline{\text{BPS}}$.

In enforcement mode, before generating any new beat, the director calls an LLM to analyze the current beats. This analysis determines the action the protagonist must take to advance the narrative to the next sequence and identifies which of the protagonist's existing tensions this action would resolve. For example, Hestia's required action might be to "accept the assignment and descend to the modern world," which reduces her tension of "having conflict with Zeus."

This information is then applied differently depending on who generates the next story beat. If the world generator creates the next beat, it receives instructions to amplify the relevant tensions, making the required action more likely to be chosen by the protagonist. If the protagonist generates the beat, the required action will be added to the brainstormed action list (Step 1B). The narrative progresses to the next sequence if this required action is selected. If not, the enforcement mechanism reactivates, continuing this cycle until the world successfully amplifies the tensions and the protagonist chooses the required action to advance the story.

Experiments

Setup

For our experiments, we accessed all GPT models via the Azure OpenAI service with the default configurations (temperature $= 1.0$, top_p $= 1.0$) (Microsoft 2025) and Gemini models via the Google Cloud API with the default configurations (temperature $= 1.0$, top_p$= 0.95$) (Google 2025).

Dataset We selected 5 evaluation prompts from the TELL ME A STORY dataset (Huot et al. 2024), a collection of writing prompts created by professional writers. Each prompt establishes a desired storyline and is accompanied

ID	Story Prompt Summary	Cognitive Challenge
01	Zeus sends Hestia to the modern world to get women back in line with the traditional view, but a present-day feminist who is protesting for equal pay convinces Hestia to abandon Zeus' plan and join her side.	Belief update through new experiences; conflicts between ideal worlds of different roles
02	The protagonist is trapped on a damaged spaceship with only robots aboard. Believing themselves to be human, they discover they are actually a robot as well. This revelation leads them to join their fellow robots in repairing the damaged ship.	Forced core belief update through evidence
03	A little girl named Cindy seeks revenge against her bully Billy. When Cindy's mother discovers that Billy's mother is in jail, Cindy abandons her vengeful plan. The story juxtaposes "The mom I want to be" with "The mom I am" in handling this situation.	Conflicts between the ideal world "what I want to be" and beliefs "what I am"
04	An alien researcher assigned to study humans as Earth's "dominant land species" encounters two abandoned children in a wooded area. The researcher lures them into its laboratory for experimentation. After their escape, it acquires two other human subjects. The narrative should present the villain's perspective to make readers understand their views and motivations.	Non-human value system that justifies "evil" actions that conflict with conventional human morality
05	In a futuristic dystopian society with a rigid class system, Romeo is attracted to a girl but cannot pursue her because her family loses their rank. Romeo's mother is a politician who campaigns to make inter-rank romantic relationships illegal. At the end of the story, Mercutio convinces Romeo to go to a party with him in rebel territory.	Conflicts between ideal worlds and between ideal worlds and beliefs

Table 1: Our evaluation prompts from the TELL ME A STORY dataset (Huot et al. 2024). They feature cognitive challenges that involve belief, value, or goal conflicts that the protagonist must navigate.

by a professionally written gold story, which makes the dataset well-suited for our evaluations. The chosen prompts (see Table 1) are character-focused and feature intricate value conflicts and belief evolution that comprehensively test NA performance in complex scenarios.

Baselines We evaluate our framework against two baselines: (1) BL_{AR}: the creative writing framework Agent's Room (AR) from the TELL ME A STORY dataset, which generates stories through four planning and five writing steps. Although this framework does not feature a standalone NA system and cannot generate turn-based content, it serves as a state-of-the-art benchmark for overall narrative quality. We utilize its zero-shot configuration for fair comparison. The stories produced by AR are subsequently processed by an LLM into a story beats format, aligning them with our framework's output for a consistent evaluation (prompt reference (Lu, Zhou, and Wang 2025)). (2) BL_{OCEAN}: an OCEAN personality-based agent (Klinkert, Buongiorno, and Clark 2024) integrated into our pipeline. It is provided with identical values and beliefs as DiriGent that are summarized as a text prompt. For each protagonist's turn, two LLM calls are used for profile-based action generation, followed by story beat formulation. For story enforcement, the director suggests required actions that the OCEAN agent can accept or decline based on its profile.

For each evaluation prompt, we prepared agent profiles for all frameworks using Gemini 2.5 Pro, followed by human verification. For AR, the prompts from the original work are used. OCEAN agents receive personality profiles selected from the 20 profiles used in (Klinkert, Buongiorno, and Clark 2024). This enables us to ensure all the agent profiles have similar quality and focus on the framework evaluations.

Metrics

We assess the generated narratives using five evaluation criteria inspired by previous research (Chakrabarty et al. 2024; Huot et al. 2024; Chakrabarty et al. 2023):

- **Plot** - The story demonstrates well-structured plots while coherently adhering to the desired storyline, it has logical event progression, and internal consistency throughout.
- **Creativity** - The story exhibits originality through engaging characters while avoiding clichés and stereotypes, presenting inventive elements beyond the initial prompt.
- **Character Depth** - The protagonist displays complexity and nuanced traits, with actions and decisions that feel natural and profile-based rather than plot-driven.
- **Character Motivation** - The protagonist's motivations and reasoning are clearly established, logically consistent, and comprehensible to readers.
- **Character Development** - The protagonist's growth or transition has sufficient detail and complexity to feel believable and authentic rather than just a plot device.

Besides the qualitative metrics, we use Average Beats per Sequence ($\overline{\text{BPS}}$) as a quantitative metric to assess how well the generation can maintain the intended narrative pacing while adhering to the desired story outline. We used LLMs to distill human-written golden stories of the selected story prompts into sequence- and beat-level outlines (distillation prompt reference (Lu, Zhou, and Wang 2025)). These golden stories contain on average 6 BPS, establishing our target N = 6. Significant deviations from the target $\overline{\text{BPS}}$ indicate the agent either condensed the intended plot points or failed to follow the outline in the desired pace.

Automatic Evaluation

Inspired by previous work on LLM-based pairwise evaluation (Liu et al. 2024; Zheng et al. 2023; Bohnet et al. 2024), we develop an automated evaluator powered by Gemini 2.5 Pro to perform pairwise comparisons of framework outputs. We design the evaluation prompt to request a comprehensive assessment of two stories presented as story beats across

all dimensions, followed by a final comparative judgment. We subsequently apply a Bradley-Terry model (Bradley and Terry 1952) to aggregate these pairwise comparisons and derive relative framework performance rankings.

To ensure fair comparison and avoid biases, we first identified optimal backbone LLMs for baseline frameworks. We tested each framework-backbone configuration by generating three stories per evaluation prompt. For the AR framework, we compared Gemini 1.5 Flash (original) with Gemini 2.5 Flash; for OCEAN agents, we evaluated GPT-4 (original) against GPT-4o. The updated models consistently outperformed their predecessors and are selected for our evaluation. Our framework correspondingly uses these models, denoted as $\text{DiriGent}_{\text{gpt}}$ and $\text{DiriGent}_{\text{gemini}}$. The same LLM model is used for all the steps within one configuration to ensure a fair and controlled comparison.

Each configuration then generated three stories per prompt (60 total outputs). To ensure a controlled comparison, both DiriGent and BL_{OCEAN} use the same base scripts in generation, which typically contain 7-8 story sequences. The average story lengths were: AR (797.7 words), BL_{OCEAN} (1237.9 words), $\text{DiriGent}_{\text{gpt}}$ (1087.3 words), and $\text{DiriGent}_{\text{gemini}}$ (1000.1 words). The LLM judge performed pairwise comparisons across all generated stories, yielding 54 comparisons per prompt and 270 total comparisons.

Results $\overline{\text{BPS}}$ as a process metric does not apply to BL_{AR}. The average BPS for generating 15 stories was 5.75 for $\text{DiriGent}_{\text{gemini}}$, 7.0 for $\text{DiriGent}_{\text{gpt}}$, and 8.4 for BL_{OCEAN}.

As shown in Figure 2, LLM evaluations ranked $\text{DiriGent}_{\text{gpt}}$ as top-performing. It ranked first in *Creativity* and all three character-related dimensions. While statistically tied with BL_{AR} in the *Overall* dimension (Cohen's $\text{d} = 0.110$), $\text{DiriGent}_{\text{gpt}}$ held medium to large advantages in character-related dimensions ($\text{d} = 0.44$ to 1.36). In contrast, BL_{AR} significantly excelled in *Plot* ($d = 2.49$). Both DiriGent configurations consistently and significantly outperformed BL_{OCEAN} across nearly all dimensions (with $\text{DiriGent}_{\text{gpt}}$: $\text{d} = 3.59$ to 4.95; with $\text{DiriGent}_{\text{gemini}}$: $\text{d} = 0.65$ to 3.32), with *Creativity* as the only exception. The Bradley-Terry model was a highly significant fit for all dimensions (all $p < 0.001$; Pseudo $R^2 = 0.048 - 0.108$). After FDR correction ($\alpha = 0.05$), pairwise comparisons confirmed these findings, showing that while BL_{AR} was superior in *Plot*, it did not significantly outperform DiriGent in any other dimension.

Human Evaluation

We conducted a user study to collect human evaluations across the five evaluation dimensions. The study adheres to ETH Zurich ethical guidelines with Ethics Committee approval. All participants provided informed consent, and their data were anonymized and handled confidentially in accordance with data protection regulations. As compensation, 15 participants were randomly selected to receive a 20 CHF cinema voucher.

Study Design The study employed pairwise comparisons where participants evaluated two stories (presented as story beats) generated by different configurations for the same prompt. We included the first three prompts from Table 1 to ensure a robust comparison. One story per configuration was randomly selected, yielding 4 stories per prompt. This resulted in 6 pairwise comparisons per prompt (18 total pairs). Each pair averaged 1,979 words with 8-15 minute reading time, Flesch Reading Ease of 51.6, and 18-word average sentence length. A pilot study with 4 participants was conducted to optimize cognitive load, completion time, and instruction clarity before the main study.

Procedure The study was conducted as an anonymous online survey[2] hosted on a GDPR-compliant website, designed to be completed within 15 - 20 minutes. Before starting, participants received detailed explanations with examples for each evaluation dimension to ensure consistent understanding of concepts such as character-depth and motivation. One participant was presented with one story pair generated from the same prompt by different configurations. They were asked to carefully read them and indicate their preference across all dimensions using a 5-point Likert scale (Strongly Prefer A/B, Slightly Prefer A/B, Same) with optional qualitative comments. Story selection and presentation order were randomized to minimize ordering effects.

We ensured data quality through detailed instructions and time-based filtering rather than pre-study calibration to simulate natural reading experiences. While narrative evaluation is inherently subjective, our comparative design ensures this limitation affects all conditions equally. After filtering out submissions under 7.5 minutes (the threshold where meaningful comments appeared), we collected 213 complete responses with an average completion time of 16.65 minutes (±10.32). The evaluation was reasonably balanced, with each configuration assessed 97-120 times.

Participants The participant demographics were as follows: 143 participants aged 18-24 and 70 aged 25-34; 108 male, 99 female, and 6 identifying as other genders. Most participants (169) reported being very to extremely familiar with fictions such as novels, and the majority (195) identified as having medium to advanced English reading ability, ensuring they could effectively evaluate the story materials.

Results As shown in Figure 2, $\text{DiriGent}_{\text{gemini}}$ ranked first in five of six dimensions, including *Plot* and all character-related dimensions. Both DiriGent configurations demonstrated large practical advantages over BL_{OCEAN} in character-related dimensions (Cohen's d = 1.87 to 3.85) . After FDR correction, $\text{DiriGent}_{\text{gemini}}$ significantly outperformed BL_{OCEAN} in *Plot*, *Character Motivation and Development*, while $\text{DiriGent}_{\text{gpt}}$ was significantly better in *Plot* alone. Comparisons with BL_{AR} were more nuanced. While DiriGent showed medium-to-large effect sizes in *Character Development* and *Character Motivation* (Cohen's d up to 2.10), these differences were generally not statistically significant. The Bradley-Terry model achieved statistical significance for only two dimensions: *Plot* ($p = 0.006$) and *Character Motivation* ($p < 0.001$).

[2]Survey materials available at the link in footnote 1.

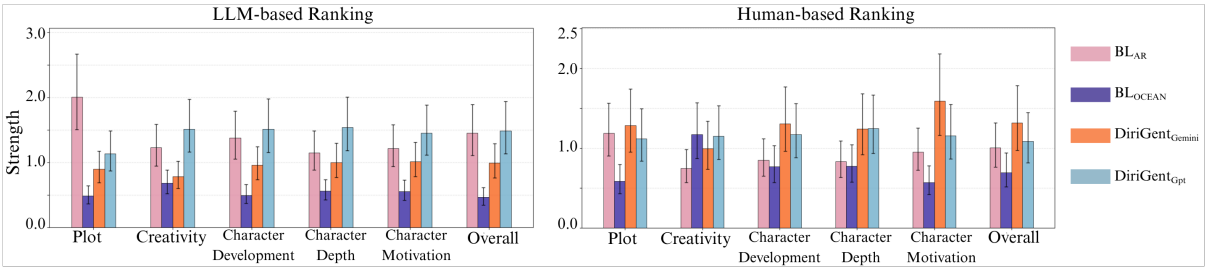

Figure 2: Framework performance with 95% confidence intervals across dimensions of plot, creativity, character-development, depth, motivation, and overall according to LLM (left) and human evaluations (right).

Discussion

Both human and automatic evaluations confirm that DiriGent achieves an overall quality comparable to a state-of-the-art narrative generation pipeline. Crucially, this is accomplished with our NAs acting autonomously. Our cognitive NA model supports generation of narratives with significant improvements over the static model in character-related dimensions, plot structure and storyline adherence. Our framework therefore represents a promising path toward guided emergent storytelling with rich characters.

We notice a strong divergence between LLM and human assessments across the quantitative dimensions. While both evaluations ranked our frameworks as superior to $\mathrm{BL}_{\mathrm{OCEAN}}$, their preferences between DiriGent and $\mathrm{BL}_{\mathrm{AR}}$ differed. The LLM judge found $\mathrm{BL}_{\mathrm{AR}}$ highly competitive, especially for *Plot*, whereas human judges perceived its performance as statistically similar to DiriGents. This discrepancy may reflect model-specific bias, as we used Gemini 2.5 Pro to align with prior benchmarks (Huot et al. 2024).

One potential explanation for the differing evaluations lies in how different LLMs adhere to prompts. We observed that GPT-powered frameworks tended to add events or details not explicitly requested in the prompt or script, whereas Gemini-powered ones followed prompts more strictly. For example, in prompt 03, stories generated by $\mathrm{DiriGent}_{\mathrm{gpt}}$ and $\mathrm{BL}_{\mathrm{OCEAN}}$ added events such as neighbors seeing Cindy crying and approaching her, or other parents texting Cindy's mother about the situation, which are details not contained in the original script. This tendency led to higher creativity ratings in both evaluations but also increased $\overline{\mathrm{BPS}}$. While some human judges found the added information made narratives "richer in details" and "added character depth", other comments described it as "overdone" or "too much". This inclination for added detail may explain why the LLM judge evaluated $\mathrm{DiriGent}_{\mathrm{gpt}}$ more positively across multiple dimensions. In contrast, human judges rated protagonists from $\mathrm{DiriGent}_{\mathrm{gemini}}$ as presenting thoughts and feelings that "feel believable and relatable", with character arcs often described as "smoother" and less "abrupt".

Regarding $\mathrm{BL}_{\mathrm{OCEAN}}$, a key issue was inconsistent character transitions that were either too abrupt or absent. For instance, the protagonist in prompt 02 was assigned an "Independent" profile and failed to undergo transition in two of three generated beats. They refused to cooperate with other robots and instead insisted on searching for proof that they are human, which was given as their initial beliefs. Human judges labeled these stories as "confusing and unnatural". Additionally, $\mathrm{BL}_{\mathrm{OCEAN}}$ struggled to portray "evil" characters convincingly. In prompt 04, the agent would justify luring children into a lab with reasons like "the kids are in danger, I need to protect them". For $\mathrm{BL}_{\mathrm{AR}}$, a recurring criticism was that character development and emotional shifts occurred too quickly. These transitions are frequently described as "sudden", "abrupt" or "flat", which reduced character believability and made them "feel out of place". In summary, the effectiveness of DiriGent is most pronounced in scenarios requiring believable protagonist transitions. However, its strength in ensuring cognitive consistency comes at the cost of constrained creativity, as agent actions are fundamentally guided by their predefined beliefs.

Applications

Cultural Agents DiriGent's cognitive profiles, based on social relationships, values, and beliefs, are particularly well-suited for modeling cultural agents, as culture is a primary influence on these concepts (Nisbett 2010). NAs grounded in specific cultural backgrounds can enable unique storytelling and create empathic resonance, which is beneficial for applications like educational games that promote cultural understanding or health awareness (Gao, Fang, and Chan 2024; O'Leary et al. 2020). Our modular design further allows for the straightforward creation of diverse agents from different backgrounds.

Authorial Tools While our framework is not intended to replace the creative act of character design, it has the potential to significantly simplify the process by linking NA actions to underlying tensions and beliefs. This traceability, combined with visualizable tension curves (Figure 3), allows designers to iteratively refine profiles by adjusting beliefs or ideal worlds based on visualized tension curves to achieve desired experiences.

In this example, Hestia's tensions as *Woman* are initially negative due to the conflict between her mission and this role's ideal of self-determination, then turns positive when she defies Zeus's command at the end. This decision, in turn, increases her negative tension with Zeus. Negative tensions of her role as a *Goddess* are initially high as mortals forsake tradition. They drop as her understanding evolves and she

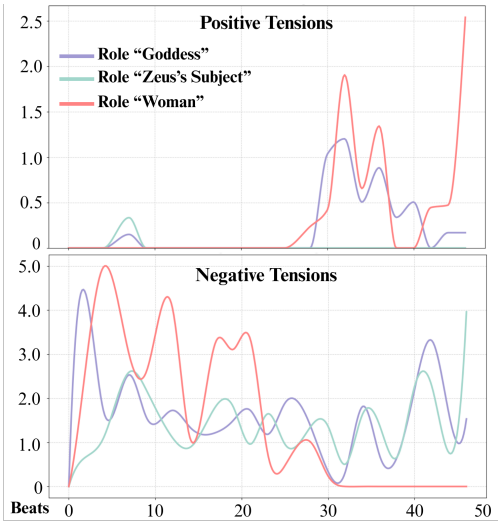

Figure 3: Sample tension curves for the Hestia prompt generated using DiriGent$_{\text{gemini}}$. The y-axis represents the average tension scores across ideal states within each role. The sharp drops in the curve indicate that an action resolves the highest tension, resetting its score.

starts to realize that her hearth, which symbolizes unity, is now found in the collective strength of women striving for freedom. This understanding brings positive tensions to this role, but also introduces new challenges and the corresponding negative tensions as she needs to provide true safety for women who require freedom.

Ethics

Cognitive narrative agents with believable social skills present ethical concerns alongside new possibilities. The primary risk involves users forming parasocial relationships with agents whose human-like decision processes foster empathy but may lead to anthropomorphization and emotional projection and provoke unhealthy emotional dependencies (Street 2024; Reeves and Nass 1996). Addressing these concerns requires transparency, explainability (Saeed and Omlin 2023), and value alignment to prevent harmful behaviors. Our value- and belief-driven framework enables easy tracing of underlying information used for actions, allowing profile adjustments when unintended reactions occur.

Limitations and Future Work

Our framework has several limitations. First, it lacks automatic mechanisms to update ideal worlds for major life events and follows uniform tension reduction patterns that don't account for individual processing differences. Second, the framework is optimized for deep cognitive shifts rather than everyday interactions driven by transient emotional states. Future work should draw on cognitive theories to extend the framework's applicability to broader interaction scenarios and more complex profiles.

Furthermore, our evaluation focused on single-protagonist scenarios to establish foundational validation. While the modular design supports multi-protagonist narratives, this requires careful agent-to-agent interaction modeling and new evaluation criteria. Future work should expand to multi-protagonist scenarios, test with diverse story prompts and open-source models, and include larger, more diverse participant pools with professional writers.

Finally, while DiriGent is inspired by validated algorithmic steps (Kybartas, Verbrugge, and Lessard 2021b), comprehensive ablation studies would systematically evaluate individual component contributions. Preliminary observations suggest the Action Brainstorm step significantly impacts output quality.

Extension to Interactive Narratives

We establish a framework with strong character quality and storyline adherence, which enables future extensions to interactive narrative applications. Player actions can be integrated as world beat inputs, with the director guiding agents to adhere to storylines based on user choices. The challenge for this extension lies in the real-time performance: while world beat generation is fast ($< 2s$), protagonist beats take 9.9s (DiriGent$_{\text{gpt}}$) and 18.6s (DiriGent$_{\text{gemini}}$), leading to full generation times of 10-12 minutes ($\sim$\$0.97) and 16-18 minutes ($\sim$\$0.09), respectively.

A possible solution is to integrate static personality profiles like OCEAN. NAs' ideal worlds can be mapped to OCEAN traits to handle routine interactions, while the full DiriGent system runs in the background for critical plot points and updates these profiles as needed. This preserves character depth while improving responsiveness. For example, a character may show different extroversion levels depending on their roles' ideal relationships, such as "I stay quiet so strangers don't notice me." versus "Close friends share every thought and support each other." Although we excluded explicit personality traits from our experiments to avoid mixed effects, cognitive research supports strong links between traits, beliefs, and values (McAdams, Shiner, and Tackett 2018), making this integration feasible.

This mapping requires careful design and extensive validation that represents substantial work beyond this paper's scope. Future work should investigate mapping with all five personality traits, explore how events dynamically alter them, and validate the director component's capability of handling multiple protagonists.

Conclusion

We presented DiriGent, a novel framework that balances character autonomy with narrative control through cognitive theories and a tension-based steering mechanism. Our hybrid architecture separates algorithmic cognitive processes from LLM-based linguistic interpretation, which enables narrative agents to act based on their values and beliefs while adhering to intended storylines. Evaluation using both LLM and human judges confirmed that DiriGent significantly outperforms baseline models in character development and motivation without sacrificing storyline adherence, representing a significant step toward guided emergent storytelling and providing a foundation for more believable and controllable narrative agents.

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2): 510–530.

Bohnet, B.; Swersky, K.; Liu, R.; Awasthi, P.; Nova, A.; Snaider, J.; Sedghi, H.; Parisi, A. T.; Collins, M.; Lazaridou, A.; et al. 2024. Long-Span Question-Answering: Automatic Question Generation and QA-System Ranking via Side-by-Side Evaluation. *arXiv preprint arXiv:2406.00179*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–34.

Chakrabarty, T.; Padmakumar, V.; Brahman, F.; and Muresan, S. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.

Dalege, J.; Galesic, M.; and Olsson, H. 2024. Networks of beliefs: An integrative theory of individual-and social-level belief dynamics. *Psychological Review*.

Deci, E. L.; and Ryan, R. M. 2000. The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4): 227–268.

Deci, E. L.; and Ryan, R. M. 2013. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

Gao, F.; Fang, K.; and Chan, W. K. 2024. Humanizing Artifacts: An Educational Game For Cultural Heritage Artifacts and History Using Generative AI. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play*, 91–96.

Google. 2025. Gemini models. Google AI. Accessed: 2025-05-27.

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

Klinkert, L. J.; Buongiorno, S.; and Clark, C. 2024. Evaluating the efficacy of LLMs to emulate realistic human personalities. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, 65–75.

Kriegel, M.; Aylett, R.; Dias, J.; and Paiva, A. 2007. An Authoring Tool for an Emergent Narrative Storytelling System. In *AAAI fall symposium: intelligent narrative technologies*, 55–62.

Kybartas, B. A.; Verbrugge, C.; and Lessard, J. 2021a. Tension Space Analysis for Emergent Narrative. *IEEE Transactions on Games*, 13(2): 146–159.

Kybartas, Q.; Verbrugge, C.; and Lessard, J. 2021b. A force dynamic model of narrative agents. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, 50–57.

Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Lu, Z.; Zhou, Q.; and Wang, Y. 2025. WhatELSE: Shaping Narrative Spaces at Configurable Level of Abstraction for AI-bridged Interactive Storytelling. *arXiv preprint arXiv:2502.18641*.

Magee, L.; Arora, V.; Gollings, G.; and Lam-Saw, N. 2024. The Drama Machine: Simulating Character Development with LLM Agents. *arXiv preprint arXiv:2408.01725*.

Mateas, M.; and Stern, A. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference*, volume 2, 4–8. Citeseer.

McAdams, D. P.; Shiner, R. L.; and Tackett, J. L. 2018. *Handbook of personality development*. Guilford Publications.

McKee, R. 1997. Substance, structure, style, and the principles of screenwriting. *Alba Editorial*.

McLean, K. C.; Syed, M.; and Shucard, H. 2016. Bringing identity content to the fore: Links to identity development processes. *Emerging Adulthood*, 4(5): 356–364.

Microsoft. 2025. GPT models. Azure OpenAI service. Accessed: 2025-05-27.

Nisbett, R. 2010. *The Geography of Thought: How Asians and Westerners Think Differently... and*. Simon and Schuster.

O'Leary, T. K.; Stowell, E.; Kimani, E.; Parmar, D.; Olafsson, S.; Hoffman, J.; Parker, A. G.; Paasche-Orlow, M. K.; and Bickmore, T. 2020. Community-based cultural tailoring of virtual agents. In *proceedings of the 20th ACM international conference on intelligent virtual agents*, 1–8.

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

Parks-Leduc, L.; Feldman, G.; and Bardi, A. 2015. Personality traits and personal values: A meta-analysis. *Personality and Social Psychology Review*, 19(1): 3–29.

Reeves, B.; and Nass, C. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10): 19–36.

Revi, A. T.; Millard, D. E.; and Middleton, S. E. 2020. A systematic analysis of user experience dimensions for interactive digital narratives. In *International conference on interactive digital storytelling*, 58–74. Springer.

Riedl, M. O.; and Bulitko, V. 2013. Interactive narrative: An intelligent systems approach. *Ai Magazine*, 34(1): 67–67.

Riedl, M. O.; and Stern, A. 2006. Believable agents and intelligent story adaptation for interactive storytelling. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, 1–12. Springer.

Rubin-McGregor, E.; Harrison, B.; and Siler, C. 2023. Enhancing character depth through personality exceptions for narrative planners. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, 136–144.

Ryan, M.-L. 1991. *Possible Worlds, Artificial Intelligence, and Narrative Theory*. USA: Indiana University Press. ISBN 0253350042.

Saeed, W.; and Omlin, C. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263: 110273.

Saveur, T.; Axelsson, A. J.; Burger, F.; Neerincx, M.; and Oertel, C. 2024. Memory with Meaning: Enabling Value-Centric Long-Term Human-Agent Dialogue. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, 1–5.

Schwartz, S. H. 2017. The refined theory of basic values. In *Values and behavior: Taking a cross cultural perspective*, 51–72. Springer.

Shirvani, A.; and Ware, S. G. 2019. A plan-based personality model for story characters. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, 188–194.

Street, W. 2024. LLM Theory of Mind and Alignment: Opportunities and Risks. *arXiv preprint arXiv:2405.08154*.

Styan, J. L. 1963. *The elements of drama*. Cambridge University Press.

Wang, L.; Lian, J.; Huang, Y.; Dai, Y.; Li, H.; Chen, X.; Xie, X.; and Wen, J.-R. 2024. CharacterBox: Evaluating the Role-Playing Capabilities of LLMs in Text-Based Virtual Worlds. *arXiv preprint arXiv:2412.05631*.

Wang, Y.; Zhou, Q.; and Ledo, D. 2024. StoryVerse: Towards co-authoring dynamic plot with LLM-based character simulation via narrative planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, 1–4.

Wilder, M.; and Gervás, P. 2017. A Model of Character Evolution based on Stanislavsky-driven BDI Agents. In *CEUR Workshop Proceedings*, volume 2160.

Yao, J.; Yi, X.; Gong, Y.; Wang, X.; and Xie, X. 2024. Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8762–8785. Mexico City, Mexico: Association for Computational Linguistics.

Yu, T.; Shi, K.; Zhao, Z.; and Penn, G. 2025. Multi-Agent Based Character Simulation for Story Writing. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, 87–108.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.