Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2025)

CoDi: A Director-Actor Framework for Goal-Driven Interactive Story Generation with LLMs

Honggu Kim, Taewoo Yoo, Yun-Gyung Cheong
Sungkyunkwan University
Suwon, Gyeonggi, South Korea
redba11@g.skku.edu, woo990307@naver.com, aimecca@gmail.com

Abstract

Interactive storytelling offers personalized and engaging narrative experiences but poses significant authoring challenges. Our proposed framework, **CoDi**, extends the existing director-actor paradigm by enhancing the director agent's control capabilities. Specifically, CoDi uses high-level narrative goals to facilitate adaptive storytelling, with the director agent able to introduce new events, select relevant non-player characters (NPCs), and explicitly describe narrative outcomes. Comparative evaluations have demonstrated CoDi's competitive narrative quality, highlighting its effectiveness in balancing structural control and flexible agent behaviors, thus underscoring its potential as a foundation for scalable interactive storytelling systems. The code is publicly available on Github.

Code — https://github.com/Speeditidious/CoDi

Introduction

Interactive storytelling has emerged as a critical medium for engaging audiences by actively involving them in shaping narrative outcomes. Unlike traditional linear narratives, interactive stories offer personalized experiences (Page 1999; Thue et al. 2007) but require significant authoring effort (Jones and Millard 2024). To address this concern, automatic story generation has been extensively studied by automating some stages in the story generation pipeline (Alabdulkarim, Li, and Peng 2021). Additionally, recent advancements in Large Language Models (LLMs) further enable real-time, interactive generation, exemplified by frameworks such as Dramatron (Mirowski et al. 2023), which allow user interaction throughout the story generation process.

Among the various automatic story generation methods, the director-actor framework provides rich opportunities for user interaction. In this framework, a director agent orchestrates actor agents through role-playing instructions. The paradigm naturally extends to interactive storytelling by allowing users to control an agent or intervene directly in simulations. For example, IBSEN (Han et al. 2024) replaces an actor agent with a human player in a game environment.

However, traditional director-actor frameworks typically focus on generating a fixed plot or a sequence of scenes, which the actor agents are then expected to follow. While this approach facilitates stable narrative simulation, it poses challenges in interactive contexts where users may alter the narrative direction unpredictably, thereby rendering the pre-established plot invalid. Figure 1 illustrates two additional key challenges inherent to director-actor based story generation in natural language environments. First, certain high-level narrative goals—referring to overarching story-level objectives such as resolving a central conflict—may remain unfulfilled without explicit interventions. Second, unlike video game environments where state changes are visually rendered, natural language-based simulations must rely entirely on textual descriptions. When these changes are insufficiently verbalized, it can compromise narrative coherence and diminish reader immersion.

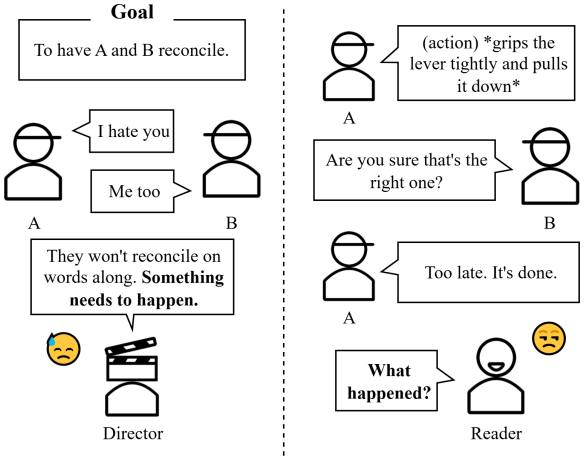

Figure 1: **Challenges in director-actor based story generation within natural language environments.** There are two key limitations: (1) Certain high-level narrative goals are difficult to fulfill without explicit intervention; and (2) changes in the environment are not visually observable unless described, which may disrupt reader immersion if under-explained.

We present **CoDi**, a director-actor framework that generates narratives dynamically, enabling flexible adaptation to interactive storytelling. CoDi guides story development

using high-level narrative goals instead of fixed endings or predefined scene sequences, maintaining structured coherence while allowing user-driven changes. CoDi addresses the challenges shown in Figure 1 by equipping the director agent with two core capabilities: (1) introducing new events or unspecified characters to steer narrative progression, and (2) verbalizing state changes and outcomes of agent actions to preserve coherence in text-based environments.

We conducted experiments, and the results show that CoDi won approximately 61.8%, lost 14.5%, and tied in the remaining 23.6% against a prior director-actor based framework and performed comparably to, or slightly better than human-written narratives across multiple story quality dimensions. These results highlight CoDi's ability to generate competitive narrative quality, demonstrating its potential as a foundational framework for interactive storytelling.

Our key contributions are as follows:

- We introduce CoDi, a director-actor simulation framework that pursues high-level narrative goals through coordinated agent interactions, enabling more flexible narrative flow than traditional scene-based methods.
- We design control instruction types for natural language environments and validate their impact through ablation studies.
- We conduct a comparative evaluation of story quality against baselines, including a prior actor-director simulation framework and human-written narratives, demonstrating CoDi's competitive performance and potential as a foundation for interactive story generation.

Related Work

Rule-Based and Goal-Driven Story Generation

Prior work in interactive narrative generation frequently leveraged rule-based and goal-driven methods. Classic systems such as TALE-SPIN (Meehan 1977) employed explicit symbolic rules to drive narrative progression by modeling characters' immediate decisions. Similarly, IPOCL (Riedl and Young 2010) integrated intentional planning to ensure coherent character behaviors aligning with narrative goals. More recently, Sabre (Ware and Siler 2021) improved goal-driven story planning by incorporate a deep theory of mind, enabling characters to act according to individual motivations rather than global directives. This narrative planner finds characters' action sequences that are not only aligned with overarching author goals but also grounded in individual characters' internal goals via utility functions that reflect each character's unique preferences and beliefs.

This goal-driven approach offers greater flexibility in narrative progression compared to approaches that follow fixed, predefined outlines. However, traditional rule-based systems operate within a predefined set of elements—such as characters, actions, and world states—which inherently limits their adaptability. In this study, we investigate how the core principles of Sabre can be extended into a more dynamic and interactive storytelling framework by leveraging the expressive and generalizable capabilities of Pretrained LLMs.

LM-Based Automatic Story Generation

The advent of large-scale neural language models significantly impacted automated narrative generation. Early neural approaches, including Neural Story Generator (Roemmele and Gordon 2018) and hierarchical neural story generators (Fan, Lewis, and Dauphin 2018), leveraged recurrent architectures to produce coherent short-form narratives. The introduction of transformer-based architectures facilitated long-range context modeling, exemplified by Plan-and-Write (Yao et al. 2019), which generated event summaries before expanding into full narratives. Further improvements in coherence were realized by DOC (Yang et al. 2023), which employed detailed outlines for paragraph-level narrative control. Beyond structural planning, narrative generation has also evolved to emphasize specific storytelling elements. For example, CNGCI (Song et al. 2024) enhances narrative tension by explicitly modeling conflict as a central component during generation through the inference of infeasible goals. Additionally, the Agent's Room framework (Huot et al. 2024) generated narrative with enhanced coherence in an end-to-end manner using multi-agent interactions.

Although these approaches can generate long and coherent stories, they do not model individual characters as distinct agents, in particular, they lack support for character-level agent modeling and cannot incorporate LLMs fine-tuned to act as individual characters, such as CharacterLLM (Shao et al. 2023). As a result, it is difficult for them to maintain character consistency. Moreover, they struggle with scenarios that involve numerous characters or extensive character profiles—including facets such as moral dispositions (Bae et al. 2025)—since such information may exceed the context-window limitations of current LLMs. Consequently, maintaining character consistency and simulating rich multi-agent interactions remain significant challenges for contemporary automatic story generation systems.

Multi-Agent Narrative Simulation

Multi-agent frameworks emerged to address interactive narrative generation by explicitly modeling character agency. HoLLMwood (Chen et al. 2024) utilized a writer-agent to produce hierarchical scene structures, subsequently interpreted by character-specific agents performing detailed actions. StoryVerse (Wang, Zhou, and Ledo 2024) similarly adopted director-actor paradigms, focusing on predefined scene-level goals executed within game environments. CoSER (Wang et al. 2025) further coordinated role-based simulations to improve human-like character interactions and fidelity.

While these frameworks offer notable narrative complexity, they commonly rely on narrowly defined plots or scene-level goals, which limits their adaptability to interactive or evolving narrative trajectories. Beyond this structural limitation, they also frequently struggle to reconcile character-driven actions with overarching narrative objectives, and generally overlook the distinctive challenge posed by textual environments that lack explicit visual cues. Our proposed framework, CoDi, seeks to bridge these gaps by empowering the director agent with capabilities to dynamically

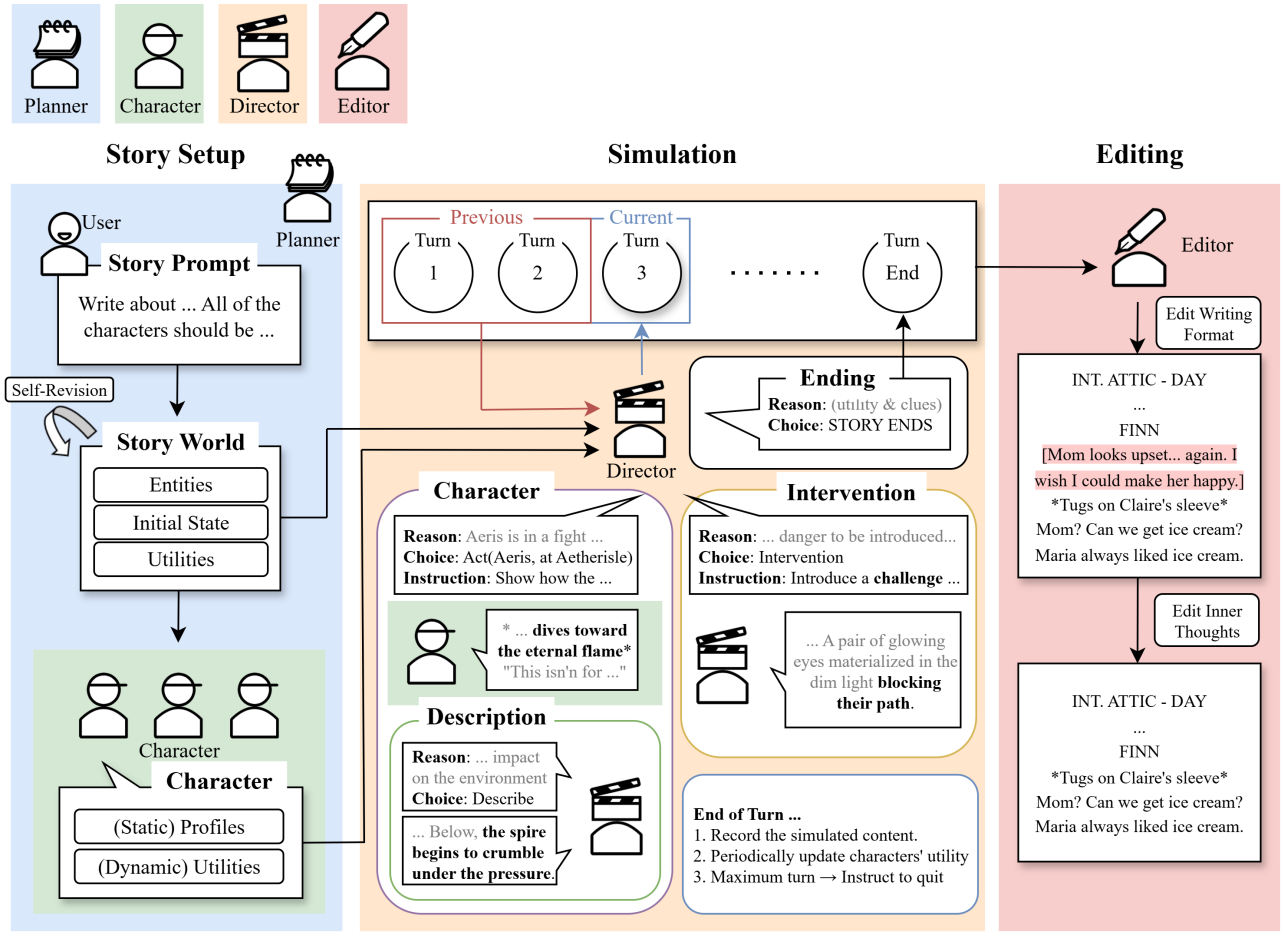

Figure 2: **Overall framework of CoDi.** CoDi consists of three phases: Story Setup, Simulation, and Editing. In the Story Setup phase, the planner agent initializes the story world and character agents. In the Simulation phase, the director agent issues instructions in a turn-based manner to advance the narrative, referencing the current story context. Lastly, in the Editing phase, the editor agent edits the simulated result into a readable format such as script or novel.

introduce narrative events, non-player characters, and explicit descriptions of story states. By maintaining high-level narrative goals and providing explicit interventions, CoDi achieves a balance between structural coherence and flexible character autonomy, facilitating scalable, interactive storytelling suited for dynamic narrative contexts.

CoDi

This section describes our multi-agent framework for story generation, which consists of four specialized agents:

- **Planner Agent.** Initializes the story by generating character profiles and high-level narrative goals.
- **Director Agent.** Guides the narrative progression by issuing high-level instructions consistent with the predefined goals.
- **Character Agent.** Enacts character-specific behaviors and dialogue in response to the director agent's instructions.
- **Editor Agent.** Post-processes the generated content into screenplay format to enhance coherence and readability.

We use `gpt-4o-2024-11-20`[1] (Achiam et al. 2023) as the backbone model for the planner agent, and `gemini-2.0-flash`[2] (Pichai, Hassabis, and Kavukcuoglu 2024) for the director, character, and editor agents.

Figure 2 presents the overall story generation process of CoDi, which consists of three main phases: (1) Story Setup, (2) Simulation, and (3) Editing.

Stage 1: Story Setup

The Story Setup phase establishes the initial context for narrative generation. It begins with a story prompt—a user-provided natural language paragraph outlining key narrative elements.

Story World Initialization Based on this prompt, the planner agent constructs a virtual environment and defines

[1]https://platform.openai.com/docs/models/gpt-4o
[2]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

Narrative Utility Function

utility(narrative):
Hannah successfully delivers TheLetter to Oliver. → score += 3
Oliver reads Meredith's letter and comes to terms with his feelings. → score += 4
Leo learns about the value of kindness through Hannah's actions. → score += 2
JennyStamp is kept safe or finds its rightful owner. → score += 1

Figure 3: **An Example of a narrative utility function defined by the planner agent in the Story Setup phase.** The function encodes high-level narrative goals, assigning score based on its importance. Note that it does not explicitly describe a specific ending, thereby allowing for flexible narrative progression.

high-level narrative goals. The structure of this phase draws inspiration from Sabre (Ware and Siler 2021), which incorporates a deep theory of mind into its narrative planning, which character agents are guided only by their individual profiles and goals without reference to global narrative goals. The story world consists of three core components: entities, initial state, and utilities.

- **Entities.** The planner agent defines entities under three predefined types: characters, places, and items. To support multi-character interaction, at least three characters are generated unless the story prompt explicitly restricts the number of characters to fewer than three. Unlike Sabre, which relies on a fixed entity set, CoDi dynamically creates entities based on the prompt, allowing for greater narrative flexibility and personalization.
- **Initial State.** This component specifies the initial conditions of entities, with particular focus on character states. Rather than using structured, logic-based representations (e.g., `alive(Tom)` as in Sabre), CoDi leverages the expressive capacity of LLMs by adopting free-form natural language descriptions (e.g., "Tom is alive"). This design avoids the need to predefine state schemas or logical predicates, thereby reducing authoring effort and supporting greater variability and naturalness. This choice also aligns with recent findings suggesting LLMs perform more effectively when using natural language representations over formal logic (Liu et al. 2024).
- **Utilities.** Utility functions encode preferences over states, supporting decision-making from both narrative and character-specific perspectives. The planner initializes a global narrative utility as well as individual utilities for each character agent. This structure supports deep theory of mind by enabling character agents to act based on their own subjective goals. Figure 3 shows an illustrative example of a narrative utility function, which avoids specifying a fixed ending, thereby supporting flexible story trajectories.

The planner agent then performs up to N rounds of self-revision (N=3 in this study), iteratively evaluating and refining the initial setup to ensure consistency and coherence. Once the setup is finalized, character agents are instantiated.

Character Agent Initialization Following the story world, the planner agent instantiates character entities as autonomous character agents. Each character is assigned a narrative role—*protagonist*, *villain*, or *side character*—based on the classification framework proposed in the Better Writers Series (Black 2017, 2018, 2020).

Each character is characterized by two feature types: static features and dynamic features. The static features are intrinsic traits that remain constant throughout the narrative. The planner expands each character's description into a detailed profile, incorporating attributes such as gender, age, core strengths and flaws, behavioral tendencies, and personal history. This process draws on narrative design principles from the Write Great Fiction Series (Bell 2008).

The dynamic features represent mutable aspects of the character, such as their goals and desires. While the planner provides initial utility functions, these may not fully align with the character's profile or internal logic. To ensure coherence, each character agent re-initializes its utility function based on its unique profile, excluding external narrative-level information such as overarching narrative goals. These functions evolve over time, enabling characters to adapt their motivations and objectives as the story progresses.

Stage 2: Simulation

The Simulation phase unfolds in a turn-based manner, where each turn represents a narrative step directed by the director.

Director Generates Instruction At each turn, the director agent references the current story context and issues a high-level instruction to advance the story. The context consists of a pre-processed setup—entities, character profiles, and narrative/character utility functions—and the accumulating story progress. To avoid context-window overflow and performance degradation, character profiles are supplied in summarized form. Rather than tracking explicit world states, CoDi follows Sabre's approach (Ware and Siler 2021) by treating the evolving narrative itself as an implicit state representation, where the agent is instructed to infer the current state by tracking the narrative context. This reduces the need for explicit state tracking.

The director agent selects one of four instruction types, each designed to enhance narrative control:

- **Character.** The director selects a character to act or react when appropriate. The output follows the format `Act(Character Name, Character Location)`, accompanied by a performance instruction. The designated character agent then responds from its own perspective, using its profile and utility function to interpret and act on the instruction in a character-consistent way.
- **Description.** When a character completes an action or observes an event, the director determines whether additional narration is necessary to maintain narrative immersion. Unlike virtual environments such as video games,

Director Agent Response

Reason: The story is nearing its climax, but **there is still room for more tension and danger to be introduced**. The story's progression, ... The narrative goals that remain are: - Dr. Arthur Whitaker reconciles his love for Nikolai with the shocking revelation. - Dr. Nikolai Ivanov chooses between protecting Arthur or fulfilling his secret mission.
Choice: Intervention
Instruction: Introduce a challenge in the narrow fissure, such as an obstacle or a threat, that forces Nikolai to choose between his mission and Arthur's safety.

Narrative Outcome of Intervention

The fissure abruptly narrowed, becoming barely wide enough to squeeze through. A low, guttural growl echoed from the darkness ahead, followed by the scraping of claws on ice. **A pair of glowing eyes materialized in the dim light, blocking their path.**

Figure 4: **An Example of Director Agent Response.** The director agent selects the Intervention instruction type to increase narrative tension and advance narrative goals. In this instance, a non-player character (an arctic wolf), not defined in the story setup, is introduced as a dynamic narrative element.

where state changes are visually apparent, natural language storytelling requires explicit verbalization of outcomes (e.g., whether a locked door opens). The director includes extra descriptions when needed to help the reader understand what happened—especially for events that are not directly shown. However, if leaving out a detail is intentional—for example, to build suspense—the director may choose not to describe it on purpose.

- **Intervention.** The director may introduce spontaneous events or environmental changes (e.g., rainfall, crowd reactions) not only to enrich the story world, but also to support the achievement of narrative goals. Such interventions introduce purposeful developments that keep the story dynamic and engaging, while also providing characters with new opportunities to progress toward their objectives.
- **Ending.** When the director judges that all narrative goals have been achieved and major events have reached resolution, it ends the simulation. To ensure a coherent conclusion, the director explicitly lists each goal and provides supporting evidence from the story to demonstrate its fulfillment.

Each instruction consists of three components: (1) **Reason**, which explains the rationale behind the selection and enhances the agent's reasoning process (Wei et al. 2022); (2) **Choice**, which specifies the type of instruction; and (3) **Instruction**, which contains the actionable directive derived from the reason and choice. An example of the director agent's output is shown in Figure 4.

Response of Aeris (at Aetherisle)

[I need to channel this anger and speed into one focused strike against Solonn's control.] *eyes blazing, gathers speed, then **dives toward the eternal flame*** "This isn't for my good, Solonn! It's to stop you!"

Description by Director Agent

The wind howls as Aeris dives. Below, **the spire begins to crumble under the pressure. Solonn's spectral energy crackles.**

Figure 5: **An Example of Character Agent Response.** The response is composed of [Inner Thought], *Action/Emotion*, and "Speech". In the following turn, the director agent decided to include a description, as the action (diving toward the flame) has an impact on the environment.

Character Generates Reactions Character agents generate their responses based on the evolving story context and the instruction issued by the director agent. However, director instructions may occasionally conflict with a character's established persona. To preserve consistency, each character agent is explicitly instructed to prioritize its full character profile and utility function over the director's command when inconsistencies arise. Additionally, we append the directive 'Interpret this instruction in a way that fits your persona. Then react accordingly.' to each instruction to encourage persona-consistent behavior.

The structure of each response draws on practical narrative design principles from the Better Writers Series (Black 2020), which conceptualizes a character's 'voice' as a composite of four elements: Action, Dialogue, Thought, and Emotion. Building on this framework, character responses are composed of the three components:

- **[Inner Thought].** Introspective reflection that reveals the character's internal reasoning and personal motivations. To support a deep theory of mind, other characters do not have access to others' private thoughts.
- ***Action/Emotion*.** A description of the character's behavior and emotional state.
- **"Speech".** Dialogue articulated in the character's voice.

Each character response must include an **Inner Thought** segment to encourage the agent to prioritize its own perspective over strictly following the instruction. To enhance character autonomy, we ensured that each character could only access its own inner thoughts while being unable to observe those of others. In addition, the response must contain at least one of the following: **Action/Emotion** or **Speech**. The character agent may include both if appropriate. This format ensures that each reaction is narratively rich and internally coherent. See Figure 5 for an illustrative example.

End of Turn A turn concludes when either the director agent or a character agent provides a response. Then, this

response is appended to the story context. To reduce computational overhead, character agents update their utility functions periodically—specifically, once every 100 turns rather than at every turn. If the number of turns approaches a predefined maximum, the director agent receives an additional instruction to bring the narrative to a close.

Stage 3: Editing

While the simulated narrative is structurally complete, further refinement is required to improve readability and presentation. In this final phase, the editor agent reformats the raw output into a screenplay-style script or novel. Additionally, the agent performs post-editing focused on selectively removing inner thoughts embedded in character responses. Since these reflections appear in every character turn, refining them reduces redundancy and improves overall narrative flow.

Since the director agent dynamically determines narrative length, the output may sometimes exceed the token limit of the editor agent. To manage this, we divide the narrative into multiple chunks and process each chunk separately. To preserve coherence, each chunk is edited with the preceding one provided as context; for earlier content, a summarized version is supplied to reduce input size. In this study, given the 8,192-token output limit of the editor agent (`gemini-2.0-flash`), we set the maximum chunk size to 6,000 tokens and the minimum to 2,000 tokens to avoid exceeding the output constraint. If the final chunk falls below the minimum threshold, it is merged with the previous chunk to prevent fragmentation.

Experiments

To assess the feasibility of CoDi, we evaluate the intrinsic quality of its generated stories. We compare CoDi's outputs against two non-interactive baselines that focus solely on narrative generation quality, without supporting interactive storytelling capabilities. Given the inherent subjectivity and difficulty of evaluating long-form narratives—even for human judges—we adopt an LLM-as-judge protocol shown to align well with human preferences, following the methodology proposed in Agent's Room (Huot et al. 2024). For consistency with the original setup, we use `Gemini-1.5-Pro`[3] (Team et al. 2024) as the evaluator model.

To mitigate order bias—a known issue in LLM-based comparative evaluations where preferences may shift based on the presentation order (Koo et al. 2024)—we employ both AB and BA testing schemes. If the evaluator either rates the two stories as equal in quality or selects different stories depending on the order, we count the result as a draw.

Dataset We use the test split of the Tell Me A Story dataset[4], introduced in Agent's Room (Huot et al. 2024). This split contains 55 story prompts, each providing a high-level narrative specification, including core themes, essential elements, and narrative goals to be achieved.

[3]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-pro

[4]https://github.com/google-deepmind/tell_me_a_story

Baselines We compare CoDi against two baselines: (1) **HoLLMwood** (Chen et al. 2024), a director-actor-based story generation system in which a writer agent produces a narrative outline that is expanded into a script, followed by actor agents who refine the script through character-driven role-play. For a fair comparison, we adopt the same initial setup and backbone model (`Gemini-2.0-Flash`). (2) **Human**, a collection of human-written reference stories provided in the Tell Me A Story dataset. As these narratives are written in a novel format, outputs of CoDi are converted into a novel style, ensuring that the observed result is not merely due to format bias.

Evaluation Metrics Our evaluation protocol follows the multi-dimensional comparison framework introduced in Agent's Room (Huot et al. 2024), which evaluates story quality across the following five metrics:

- **Plot.** Does the story have a coherent structure with purposeful events and no logical inconsistencies?
- **Development.** Are the characters and settings in the story well-developed with sufficient detail and context to support realism and comprehension?
- **Language Use.** Is the language varied and rich, using literary devices effectively while avoiding bland or repetitive phrasing?
- **Creativity.** Does the story feature original ideas, engaging themes, and avoid cliches or stereotypes unless used intentionally?
- **Overall.** Which story is overall preferred, considering all aspects?

All five metrics are used when comparing CoDi to the **Human** stories. However, in comparisons with **HoLLMwood**, we omit the Creativity metric, as both systems use the same story setup to limit creative divergence. Instead, we introduce two character-centric metrics—adapted from CoSER (Wang et al. 2025)—to better assess CoDi's agent-based narrative capabilities:

- **Anthropomorphism.** Do the characters behave like autonomous humans with consistent goals and realistic decision-making?
- **Character Fidelity.** Do characters act, speak, and make decisions in ways that align with their established profiles and relationships?

Results

Figure 6 presents the comparison results of story quality across baselines. In the comparison with **HoLLMwood**, CoDi consistently produced narratives of higher overall quality. The performance gap narrows slightly on character-centric metrics such as **Anthropomorphism** and **Character Fidelity**, likely because both systems employ agent-based character modeling. While HoLLMwood relies on predefined outlines and explicit performance guidance for characters, CoDi allows for greater plot flexibility. Despite this increased adaptability, CoDi outperforms HoLLMwood, underscoring its effectiveness for open-ended, dynamic story generation.

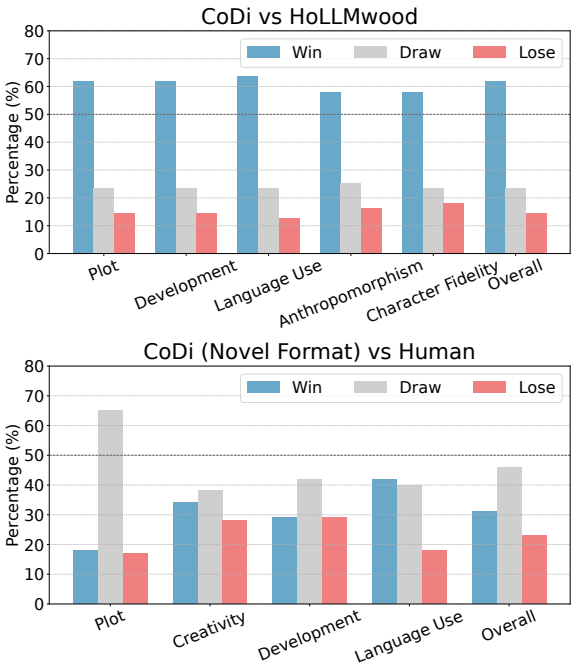

Figure 6: **Story quality comparison against the baselines.** Higher win ratio (blue bar) indicates that our framework is better. A draw is counted when the evaluator selects both stories as equally good or chooses a different one after reordering. CoDi shows competitive performance to the baselines.

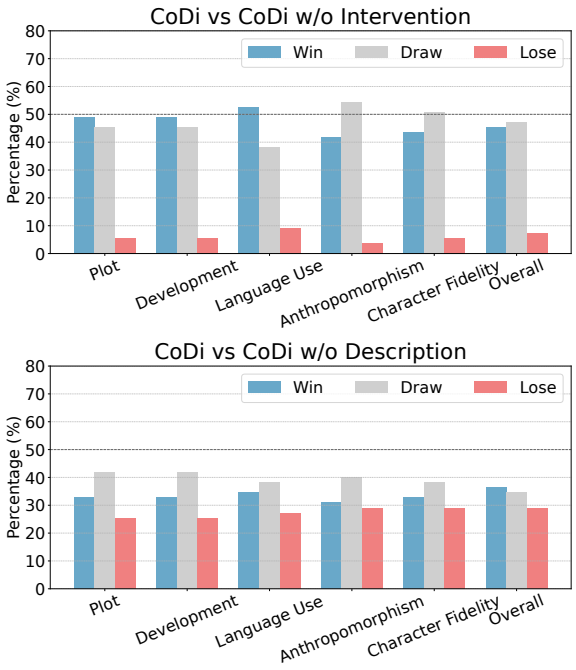

Figure 7: **Story quality comparison results when the Intervention or Description instruction type is ablated.** Ablating either instruction type negatively affected overall story quality, with the removal of Intervention having a particularly strong effect.

For a fair comparison with **Human**—a set of human-written reference stories presented in a novel-style format—we convert CoDi outputs to a novel format to mitigate format-related bias. Under this setting, CoDi demonstrates competitive performance across the story quality dimensions. Higher win rate of **Creativity** and **Language Use** indicate that our framework produced richer narratives. Notably, the **Plot** metric shows a draw rate exceeding 60% and the **Development** metric shows same win and loss rate, indicating that CoDi's generated narrative structures are comparable in story quality to those authored by humans. These findings suggest that CoDi not only advances agent-based narrative simulation beyond prior frameworks, but also approaches human-level performance in core aspects of storytelling, supporting its potential as a foundation for interactive narrative generation.

Ablation Study

CoDi introduces two instruction types—**Intervention** and **Description**—to enhance the director agent's control capabilities and address the unique demands of natural language storytelling. To evaluate the contribution of each mechanism, we perform an ablation study in which each instruction type is independently removed, and the resulting impact on story quality is measured. To reduce experimental variance, all conditions—including our full model and ablation variants—begin with the same initial turn: a paragraph describing the initial state and opening of the story.

Figure 7 presents the results. Removing **Intervention** results in a notable decline in overall story quality, with the most substantial drop observed in the **Language Use** metric. This aligns with expectations, as Intervention allows the director to introduce dynamic events and NPC reactions, which diversify vocabulary and narrative texture. Subsequent performance drops in the **Plot** and **Development** metrics further suggest that character-driven progression alone is insufficient for maintaining coherent and engaging narrative trajectories. These findings highlight the importance of balancing agent autonomy with top-down narrative control.

In contrast, the removal of **Description** yields a smaller yet measurable decline in quality. The largest drop occurs in the **Development** metric, which evaluates contextual depth. This result confirms the role of Description in reinforcing world-building and ensuring that implicit state changes are clearly conveyed in text, compensating for the lack of visual representation in natural language storytelling.

Analysis

Goal Achievement

CoDi adopts a goal-driven storytelling paradigm, which offers greater narrative flexibility compared to scene-based approaches that rely on discrete, pre-defined segments. However, this flexibility introduces additional challenges in en-

suring that narratives remain focused and consistently fulfill their intended objectives. Thus, it is critical to evaluate whether CoDi reliably achieves its defined narrative goals.

Inspired by the evaluation methodology in CoSER (Wang et al. 2025), we prompt an LLM evaluator with a list of narrative goals and ask it to assess whether each goal has been achieved. Rather than adopting CoSER's penalty-based scoring—which can be biased by narrative length—we employ a reward-based scoring scheme on a 0-1 scale: 0 for unmet goals, 0.5 for partially achieved goals, and 1 for fully achieved goals. Goal achievement serves as a more objective and quantifiable criterion.

Table 1 presents the goal achievement scores for both CoDi and its ablated variants. CoDi attains an average goal achievement rate of 81.8%, demonstrating its effectiveness in goal-driven narrative generation. As 50% reflects partial success, this result indicates that CoDi consistently fulfills its narrative objectives. Among the ablated variants, removing the **Description** instruction yields a negligible impact (81.7%), suggesting that its primary contribution lies in enhancing contextual richness rather than directly advancing goal completion. In contrast, removing **Intervention** results in a substantial performance decline (62.5%), indicating that character-driven responses alone are insufficient for reliably achieving narrative goals. This finding is particularly relevant for interactive storytelling, where real-time user-defined narrative goals require balancing agent autonomy and top-down control.

Application of Plot Structure Theory

This section explores the use of a pre-simulation human-in-the-loop mechanism to incorporate narrative theory. Numerous plot structure models exist, such as the three-act structure with fifteen beats (Snyder 2005), and no single structure is universally preferred, as audience expectations and narrative goals vary. Thus, flexible support for various structures is desirable in interactive storytelling.

To examine this capability, we conduct an experiment to assess whether CoDi can accommodate a four-part plot structure consisting of *Setup*, *Reaction*, *Attack*, and *Resolution*, as proposed by Brooks (2011). We provide the planner agent with both the initial story setup and a brief description of each structural phase. The agent then generates a set of narrative goals aligned with each part of the structure. During simulation, the system references the goals associated with the current phase to guide narrative progression. Additionally, each plot part serves as a unit of progression for updating character utility functions.

Method	Goal (Avg)
CoDi	**81.8%**
- w/o Description	81.7%
- w/o Intervention	62.5%

Table 1: **Goal achievement of CoDi (higher is better).** The best result is highlighted in **bold**. CoDi consistently meets narrative goals, with 50% indicating partial success. The ablation of Intervention notably reduced performance, underscoring its importance in narrative progression.

Method	Goal (Avg)	Theory (Avg)
CoDi w/ Theory	91.7	89.7
- w/o Description	**92.0**	**92.7**
- w/o Intervention	84.4	76.4

Table 2: **Goal achievement of CoDi with defined plot structure theory.** Providing more fine-grained narrative goals led CoDi to generate narratives with higher goal achievement.

Figure 8 presents the results of story quality comparison between two configurations of CoDi: a variant augmented with a narrative structure theory and the default version. The version incorporating the narrative structure consistently outperforms the default configuration in overall story quality. As both versions share the same underlying architecture, including agents, planning procedures, and input prompts, the comparison yields a relatively high draw rate of approximately 50%. Among all evaluation metrics, the most significant improvement is observed in the Plot category, indicating that explicit narrative structuring contributes to stronger plot development and coherence.

Table 2 summarizes the results of narrative goal achievement for the variant augmented version. Furthermore, we instruct the evaluator to assess adherence to the provided narrative structure theory. The evaluation scoring utilizes a 0-1 scale to rate fulfillment. The structured variant of CoDi achieves a higher goal achievement rate of 91.7%, suggesting that integrating narrative structure improves the ability to fulfill its narrative goals—possibly because the narrative structure segments the story into smaller, more manageable units for the director agent to oversee. The ablation analysis follows the same trend as the default setting: removing Description results in minimal change (92.0%), while removing Intervention significantly lowers performance (84.4%).

These results demonstrate that incorporating narrative structure enhances both story quality and goal achievement,

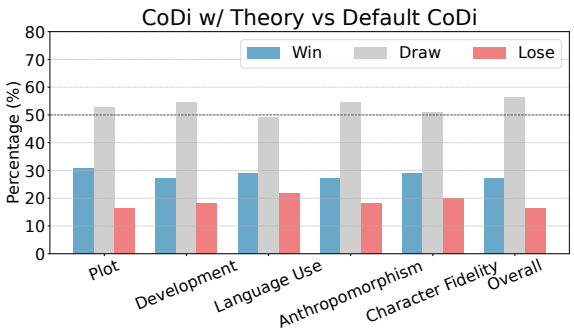

Figure 8: **Story quality comparison: variant CoDi with plot structure theory vs. default CoDi.** This indicates the potential of further improvement of story quality with user-defined narrative structure.

Method	#words	#turns	#quit
HoLLMwood	10,600	-	-
Human	1,412	-	-
CoDi	3,087	77	1
CoDi w/ Theory	8,212	206	-
- 1. Setup	1,099	29	0
- 2. Reaction	2,861	72	1
- 3. Attack	2,284	56	0
- 4. Resolution	1,968	49	2

Table 3: **Statistical Analysis.** #words and #turns denote the average number of words and turns per narrative, respectively. #quit indicates the number of narratives (out of 55) that were forced to quit due to reaching the maximum allowable number of turns, which is 200.

without compromising the flexibility of the framework, underscoring CoDi's potential as a foundation for adaptive and interactive storytelling systems.

Analysis of Narrative Length and Evaluation Bias

Table 3 reports a statistical analysis of generated narratives. The results reveal distinct patterns across systems. **HoLLMwood** produced the longest narratives on average, likely due to its writer agent occasionally generating overly detailed outlines—such as nine plot points, each with multiple subplots—despite being prompted to follow a two-level hierarchical structure. In contrast, the **Human** stories resulted in the shortest narratives, due to their novel writing format, which omits scene headings and character names.

CoDi with plot structure theory generated longer narratives than the default configuration, as expected. The addition of structure-based narrative goals increases the total number of objectives to be fulfilled during simulation, naturally extending story length. Notably, the number of narratives that reached the predefined maximum turn limit and were forced to terminate prematurely remained low (0–2 cases out of 55), suggesting that the director agent is generally effective at identifying appropriate narrative endings before the hard limit is reached.

While story length can affect evaluation results, with longer texts being favored by evaluators (Saito et al. 2023; Hu et al. 2024), our results shows that CoDi's strong performance is not simply because it generates more text. CoDi outperformed HoLLMwood, even though HoLLMwood's stories were longer. Moreover, the plot-structured version of CoDi, which produced even longer narratives, achieved the highest win ratios. These results suggest that CoDi's advantage mainly comes from the quality of its storytelling.

Limitations

CoDi has demonstrated promising results. Nevertheless, several limitations should be acknowledged. First, the evaluation relied on an LLM-as-judge framework. To mitigate potential biases, we carefully designed the evaluation process by adopting a framework with high correlation to human judgments, employing a different LLM from those used in story generation, conducting both AB and BA tests, and standardizing the story format. However, it remains unknown whether these outcomes would be replicated with human evaluators. Further validation with sufficiently large datasets and human participants is necessary in the future work.

This study focused on enhancing the director agent's ability to guide narrative development through high-level narrative goals. Experimental results indicate that CoDi generates stories that effectively achieve their intended narrative objectives, even without an explicitly planned plot structure. Although CoDi incorporates several mechanisms to increase the autonomy of character agents—for instance, dynamically updating each character's utility function over time and restricting access to other characters' internal states—the extent to which these agents can act independently, disregard the director's instructions, or engage in complex behaviors such as deception should be examined further.

Finally, CoDi uses a turn-based simulation, allowing for dynamic interaction and intervention during the story. This design can increase computational cost compared to single-shot story generation. Further research is needed to reduce computation without sacrificing narrative quality. Balancing narrative control, agent complexity, and computational efficiency remains a central challenge for making interactive storytelling more scalable.

Conclusion

We present **CoDi**, a director-actor framework for interactive storytelling that balances global narrative control with character-level autonomy. By equipping the director agent with high-level narrative goals and specialized instruction types—such as *Intervention* and *Description*—CoDi enables flexible, adaptive, and coherent story generation in natural language environments. Experimental results demonstrate that CoDi achieves competitive story quality and consistently fulfills narrative goals, outperforming prior director-actor approaches and even matching the performance of human-written stories in several aspects. We further showed that integrating plot structure theory into CoDi via a pre-simulation human-in-the-loop mechanism enhances both story coherence and goal achievement. CoDi offers a promising foundation for narrative systems that combine structured planning, agent-based simulation, and interactive flexibility. Future work will include scaling the framework to longer narratives, supporting more complex user interactions, and enhancing character autonomy for emergent storytelling.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, AI Graduate School Support Program(Sungkyunkwan University)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00357849).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alabdulkarim, A.; Li, S.; and Peng, X. 2021. Automatic Story Generation: Challenges and Attempts. In Akoury, N.; Brahman, F.; Chaturvedi, S.; Clark, E.; Iyyer, M.; and Martin, L. J., eds., *Proceedings of the Third Workshop on Narrative Understanding*, 72–83. Virtual: Association for Computational Linguistics.

Bae, S.; Cho, G.; Cheong, Y.-G.; and Li, B. 2025. CharMoral: A Character Morality Dataset for Morally Dynamic Character Analysis in Long-Form Narratives. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 8809–8818. Abu Dhabi, UAE: Association for Computational Linguistics.

Bell, J. S. 2008. *Revision And Self-Editing (Write Great Fiction)*. Write Great Fiction. Cincinnati, OH: Writer's Digest Books. ISBN 978-1582975085.

Black, S. 2017. *13 Steps to Evil: How to Craft Superbad Villains*. Better Writers Series. United Kingdom: Better Writers.

Black, S. 2018. *10 Steps to Hero: How to Craft Kick-Ass Protagonist*. Better Writers Series. United Kingdom: Better Writers.

Black, S. 2020. *8 Steps to Side Characters: How to Craft Supporting Roles with Intention, Purpose, and Power*. Better Writers Series. United Kingdom: Better Writers.

Brooks, L. 2011. *Story Engineering: Mastering the 6 Core Competencies of Successful Writing*. Cincinnati, OH: Writer's Digest Books. ISBN 9781582979984.

Chen, J.; Zhu, X.; Yang, C.; Shi, C.; Xi, Y.; Zhang, Y.; Wang, J.; Pu, J.; Feng, T.; Yang, Y.; and Zhang, R. 2024. HoLLMwood: Unleashing the Creativity of Large Language Models in Screenwriting via Role Playing. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8075–8121. Miami, Florida, USA: Association for Computational Linguistics.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898. Melbourne, Australia: Association for Computational Linguistics.

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

Hu, Z.; Song, L.; Zhang, J.; Xiao, Z.; Chen, Z.; and Xiong, H. 2024. Explaining length bias in llm-based preference evaluations. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

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*.

Jones, J. D.; and Millard, D. 2024. Experiencing The Authorial Burden. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, 78–87.

Koo, R.; Lee, M.; Raheja, V.; Park, J. I.; Kim, Z. M.; and Kang, D. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 517–545. Bangkok, Thailand: Association for Computational Linguistics.

Liu, T.; Xu, W.; Huang, W.; Zeng, Y.; Wang, J.; Wang, X.; Yang, H.; and Li, J. 2024. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. *arXiv preprint arXiv:2409.17539*.

Meehan, J. R. 1977. TALE-SPIN, An Interactive Program that Writes Stories. In *Ijcai*, volume 77, 91–98.

Mirowski, P.; Mathewson, K. W.; Pittman, J.; and Evans, R. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–34.

Page, B. 1999. Hamlet on the holodeck: The future of narrative in cyberspace. *MFS Modern Fiction Studies*, 45(2): 553–556.

Pichai, S.; Hassabis, D.; and Kavukcuoglu, K. 2024. Introducing Gemini 2.0: our new AI model for the agentic era.

Riedl, M. O.; and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39: 217–268.

Roemmele, M.; and Gordon, A. 2018. An Encoder-decoder Approach to Predicting Causal Relations in Stories. In Mitchell, M.; Huang, T.-H. K.; Ferraro, F.; and Misra, I., eds., *Proceedings of the First Workshop on Storytelling*, 50–59. New Orleans, Louisiana: Association for Computational Linguistics.

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-LLM: A Trainable Agent for Role-Playing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13153–13187. Singapore: Association for Computational Linguistics.

Snyder, B. 2005. *Save the Cat!: The Last Book on Screenwriting You'll Ever Need*. Studio City, CA: Michael Wiese Productions.

Song, Y.; Cho, G.; Kim, H.; Kim, Y.; Bae, B.-C.; and Cheong, Y.-G. 2024. A Conflict-Embedded Narrative Generation Using Commonsense Reasoning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 7744–7752. International Joint Conferences on Artificial Intelligence Organization. AI, Arts Creativity.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Thue, D.; Bulitko, V.; Spetch, M.; and Wasylishen, E. 2007. Interactive storytelling: A player modelling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 3, 43–48.

Wang, X.; Wang, H.; Zhang, Y.; Yuan, X.; Xu, R.; Huang, J.-t.; Yuan, S.; Guo, H.; Chen, J.; Wang, W.; et al. 2025. CoSER: Coordinating LLM-Based Persona Simulation of Established Roles. *arXiv preprint arXiv:2502.09082*.

Wang, Y.; Zhou, Q.; and Ledo, D. 2024. StoryVerse: Towards Co-authoring Dynamic Plot with LLM-based Character Simulation via Narrative Planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400709555.

Ware, S. G.; and Siler, C. 2021. Sabre: A narrative planner supporting intention and deep theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, 99–106.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

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

Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; and Yan, R. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7378–7385.