WhatIF: Branched Narrative Fiction Visualization for Authoring Emergent Narratives using Large Language Models

Aditi Mishra
Autodesk Research & Fujitsu Research of America
Pittsburgh, Pennsylvania, USA
amishra@fujitsu.com

Frederik Brudy
Autodesk Research
Toronto, Ontario, Canada
frederik.brudy@autodesk.com

Qian Zhou
Autodesk Research
Toronto, Ontario, Canada
qian.zhou@autodesk.com

George Fitzmaurice
Autodesk Research
Toronto, Ontario, Canada
george.fitzmaurice@autodesk.com

Fraser Anderson
Autodesk Research
Toronto, Ontario, Canada
fraser.anderson@autodesk.com

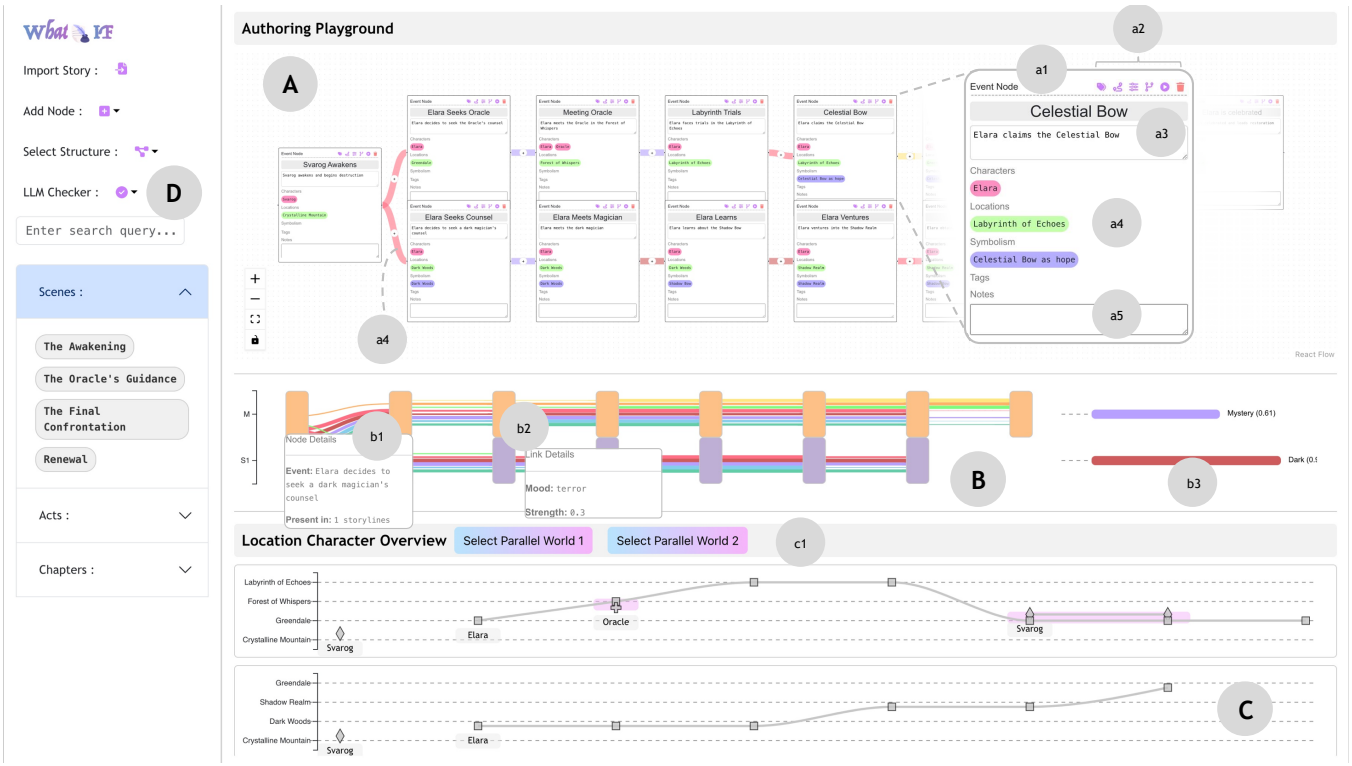

Figure 1: The WhatIF system which provides users with (A) network visualization to make and visualize narrative graphs for creative exploration, (B) detailed views of selected narrative branched based on their thematic moods, and (C) storyline visualizations of character interactions based on their locations and finally also provides (D) avenues for custom metric defined verification.

Abstract

Branched Narrative Fiction (BNF) are non-linear, text based narrative games, where the player of the game is an active participant shaping the story. Unlike linear narratives, BNF allows players to influence the direction, outcomes, and progression of the plot. A narrative fiction developer designs these branching storylines, creating a dynamic interaction between the player and the narrative which requires significant time and skill. In this work we build and investigate the use of a visual analytics tool to help narrative fiction

C&C '25, Virtual, United Kingdom

ACM ISBN 979-8-4007-1289-0/25/06
https://doi.org/10.1145/3698061.3726933

developers generate and plan these parallel worlds within a BNF. We present WhatIF, a visual analytics tool that aids BNF developers to create BNF graphs, edit the graphs, obtain recommendations, visualize differences between storylines and finally verify their BNF on custom metrics. Through a formative study (3 participants) and a user study (11 participants), we observe that WhatIF helps users plan and prototype their BNF, provides avenues to support iterative refinement of narrative and also aids in removing writer's block. Furthermore, we explore how contemporary generative AI (GenAI) tools can empower game developers to build richer and more immersive narratives.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; *Empirical studies in HCI*.

Keywords

Creativity Support, Storytelling, User Interface Design, Visualization

ACM Reference Format:

Aditi Mishra, Frederik Brudy, Qian Zhou, George Fitzmaurice, and Fraser Anderson. 2025. WhatIF: Branched Narrative Fiction Visualization for Authoring Emergent Narratives using Large Language Models. In *Creativity and Cognition (C&C '25), June 23–25, 2025, Virtual, United Kingdom.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3698061.3726933

1 Introduction

Branched Narrative Fiction (BNF) is a form of non-linear storytelling experience where the audience can influence branched storylines through their choices [57]. Each audience interacts differently, creating a unique storytelling experience for everyone. This level of engagement and agency over the narrative progression driven through their decisions and actions fosters a deeper level of involvement compared to traditional linear storytelling [26]. Early work such as Choose Your Own Adventure [3] paved the way for branched storytelling. This concept has influenced modern video games as well as mainstream media, such as the Black Mirror episode Bandersnatch [47], enabling audience to participate in the storyline.

However, authoring branched narratives requires skills and time [58], whether that involves writing fan fictions, designing indie games, or creating Dungeons and Dragons branched pathways. Authors need to plan the overarching narratives to create multiple interconnected or separate storylines, while balancing branching choices to ensure each narrative path is coherent and meaningful without redundancy. Additionally, authors need to foster audience engagement by designing decisions that feel impactful and rewarding, while preserving their authorial intent. To manage this complexity, authors use tools like Twine [70] and ChoiceScript [48] for planning and organization. However, script-based tools such as Inform7 [23] often require coding skills, creating a steep learning curve for beginners. Moreover, tools like Twine are typically used after foundational narrative elements, such as storylines and overarching structure are set, serving primarily as aids to assemble these into a playable format. Authors thus have to rely on manual, iterative processes for experimentation and testing—a time-consuming and error-prone approach.

Recently Large Language Models (LLMs) such as OpenAI GPT [2] have enabled authors to prompt their intents to generate linear and non-linear storylines [9, 17, 65]. While some focus on event based progressions, others such as Charisma.ai [13] adopt character-driven narrative development, enabling character behaviours to naturally emerge using generative agents [52]. However, LLM-based approaches make it challenging for authors to control the narrative and convey their authorial intent with simple textual prompts [31]. Moreover, recent studies suggest that LLMs may lack the creativity [12, 68] needed to generate engaging storylines, indicating the need for a mixed-initiative, human-in-the-loop approach to help authors create richer branched narrative.

These challenges led to our development of WhatIF, an interactive visualization tool designed to support authors in crafting and experimenting branched narrative fiction. By integrating an LLM, WhatIF enables authors to explore the "what-if" scenarios by transforming a simple linear story into a branched narrative, where key events become branching points, leading to diverging pathways and different outcomes. It uses a node-link interface that visually organizes these branching points, enabling authors to iterate on the storylines. The system offers detailed views for each storyline with parameters such as moods, locations, and characters. Authors can experiment with alternative branches by adjusting these parameters, prompting the system to generate and update the graph accordingly. The system also incorporates analytical tools for evaluating the branched narrative based on user-defined metrics such as narrative coherence. This allows authors to quickly experiment with different narrative possibilities, while maintaining their authorial intent. The contributions of this paper include the following: (1) an interactive visualization tool WhatIF that supports authors in crafting, experimenting, and analyzing branched narrative fictions, and (2) findings from user studies (n=14), demonstrating that WhatIF effectively enables rapid experimentation with actionable analytics insights and mood-based and location-character-based visualizations, empowering authors to quickly draft and refine branched narrative fictions.

2 Background and Related Work

This work draws on prior research in AI-assisted narrative generation, interactive visualization tools, and creative experimentation for linear narratives to inform the design of our branched narrative fiction authoring tool.

2.1 AI-assisted Branched Narrative Fiction Authoring

Various tools have been proposed to support authoring branched narratives [23, 48, 70]. They usually organize the structure in explicit branches for intuitiveness, including flowchart-like structures [70], state machine [25], and atomic storylets [32]. The creation workflow typically requires every possible event to be manually authored, which is a time-intensive and engineering-heavy process [62]. To automate this process, early computational branched narrative generation often centered on symbolic narrative planning.

This approach uses formalized plans or rules to orchestrate plot progression. Most work [6, 35, 43, 54, 58, 73, 77] focussed on classic AI planning techniques which require developers to define goals, and set possible predefined actions the characters can take and world states to generate coherent storylines using formal language. These systems attempted to preserve causal coherence by ensuring that the character decisions aligned with both narrative goals and domain constraints [58]. Despite showcasing the potential of AI in crafting complex narratives, these early systems required extensive engineering work to create a knowledge base (e.g world rules, character settings) [34]. This causes not only authorial overhead but also domain rigidity where small changes in the planning logic could potentially necessitate larger revisions of entire domain leading to scalability issues [72].

Recently Large Language Models (LLMs) have dramatically reduced the overhead of story content generation by generating text on demand due to their instruction following and semantic event tracking abilities [2, 27]. This has led to many LLM-driven works where LLMs assist in world building [19, 27, 46, 56], character development [16, 55] and plot progression [33, 46, 53, 72, 76]. Commercially available tools such as AI Dungeon [21] and Charisma.ai [13] also provide users with avenues to generate branched pathways.

However, LLM-based approaches often lack the nuanced control to preserve authorial intents as they may struggle to consistently adhere to desired narrative goals or character motivations [72]. Moreover, recent studies have found that LLMs perform significantly worse than humans in creativity tests, particularly to evoke complex emotions in storytelling [10, 12] or reduce content diversity [50]. These models produced homogenously positive storylines which lack tension with plot holes [68]. Additionally, most existing solutions concentrate on isolated aspects of branched narratives such as narrative structure [5], plot progression [13], character creation [29], or world building [18] rather than offering a holistic workflow for rapid experimentation. These reasons call for a more mixed-initiative approach, involving human-in-the-loop.

In this work, we aim to improve the controllability of generative narrative fiction by specifying their authorial intent through both storyline-based prompting and adjustable parameters—such as mood, characters, and locations. We also provide robust planning and visualization features that allow authors to structure, evaluate, and refine complex branched narratives for rapid prototyping.

2.2 Visual Analytics for Narratives

Visualizations have emerged as a critical means for analyzing and communicating narratives. Survey works by Chen et al. [14] and Tong et al.[69] provide comprehensive overviews of the role of visualization in narrative analysis. Building on these foundations, we focus on describing tools we used as sources of inspiration for WhatIF.

Conceptually, branched narratives can be considered as a graph that represents the causal relationships between events, situating it within a broader area of event sequence visualizations. Early efforts in visualizing event sequences concentrated on linear representations, laying out events along a temporal axis [4, 7, 74]. While effective for straightforward temporal data, such approaches struggle to capture the complexity in narratives that exhibit hierarchical or branching structures. To address this, subsequent systems—such as CoreFlow [40] and DecisionFlow [24] automatically extract and depict branching patterns by recursively ranking, dividing, and trimming sequences to produce interpretable tree structures. Similarly, EventThreads [28] employs tensor analysis to cluster event sequences into latent stages and evolution patterns. While WhatIF uses event node aggregation similar to prior work, our primary focus is empowering users to directly author and iterate on their narratives. Rather than relying on opaque, automated clustering, WhatIF grants direct access to individual events, letting authors manually organize them according to their creative intent.

Our location-character visualization is inspired by prior works [38, 41, 49, 66, 67, 75]. Traditional designs often place characters along the y-axis, which can obscure genuine narrative interactions and cause visual clutter due to line crossings. To address these issues, we use a layered approach that maps the y-axis to locations while representing characters as distinct paths, thereby clarifying transitions and enhancing the identification of storyline trends [8].

2.3 Creative Experimentation for Generating Linear Narratives

As noted by Wang et al. [20, 71], creative writing is inherently demanding — not only must a writer generate an overarching narrative, but they must also ensure that individual story events coherently align with that narrative. To generate coherence stories, Dramatron [45] proposes hierarchical text generation for scripts and screenplays generation. Prior work such as Talebrush [17] and StoryDiffusion [36] has demonstrated methods—using sketch-based and text-based prompting—to "flesh out" stories quickly. Similarly, FairyTailor [11] leverages multimodal inputs to generate multiple storylines for children's stories. Recently Luminate [63] supports quick experimentation to visualize story instances in a design space. While these systems rapidly generate storylines, but fall short in supporting fully branched narrative structures. In contrast, Crafting such narratives is challenging because each branch must diverge meaningfully while maintaining overall coherence—and without human guidance, the number of branches can quickly become overwhelming. WhatIF addresses this by enabling users to either manually create alternate events or use LLM-assisted branch recommendations to generate diverse, coherent storylines.

3 Design Space Exploration

To better understand the authoring process of branched narrative fictions, we performed a formative study by interviewing three branched narrative fiction (BNF) authors. From the study, we summarized the results and identified three design challenges that authors faced in their creation process. Based on the findings, we formulate five design goals to help authors exploring possible branched storylines and analyzing them in details.

3.1 Formative Interview

We conducted semi-structured Zoom interviews (1 hour each) with three branching narrative fiction (BNF) experts aged 27–36. Our participants included an institutional hobbyist (P1) and two seasoned professionals (P2 and P3) with over 12 years of BNF authoring experience—P3 also runs a Patreon for their published games

and community guidance. The interviews focused on obtaining insights from the different stages of BNF development - from the initial ideation to creative planning, iterative development, prior tool usages and challenges they encounter in the process. Through thematic coding of their responses, we identified four key design stages in the BNF creation workflow, capturing the overarching themes of developing branching narrative fictions.

3.1.1 Grounded Start. Participants mentioned starting from a grounding or a "hook", with a specific scene, location, or structure in mind. There were different "hooks". For instance, P1 envisioned a dilemma similar to the scene in the movie Titanic, where Jack is drowning while Rose is on a wooden raft. P3 started with an interesting location in mind where the story would pan out, or sometimes with a narrative structure in mind, crafting the story to fit the desired framework.

3.1.2 Iterative Expansion. Once participants had established the initial "germ of an idea", they typically proceeded to create and iterate on story branches. This process involved expanding each branch to a logical conclusion, to build out the narrative graph (Figure 2). P2 described an iterative process of rewriting branches when they encountered dead ends or could not satisfactorily complete a storyline. This iterative approach to building and refining the graph contributed to the development of more complex and engaging narratives. While each branch aligned with the overarching narrative they aimed to create, each storyline within the branched narrative fiction remained largely independent.

3.1.3 Constant Verification. Both P2 and P3 emphasized the importance of maintaining intra-storyline consistency to ensure audience satisfaction based on choices provided. For instance, if a character is killed off in a particular storyline, subsequent events should not feature that character. All participants highlighted the importance of maintaining narrative consistency. P2 and P3 specifically mentioned performing various verification tasks to ensure the consistency within each storyline, particularly after adding events or expanding branches. These checks are vital to preserving coherence and ensuring audience immersion.

3.1.4 Branched Narrative Parametrization. While branched narrative fictions can be characterized by various parameters, such as characters, locations, events, moods, and narrative structures [42], we observed that participants tended to focus on only a subset of these when beginning their BNF development. For example, P1 and P2 primarily engaged with events, characters, locations, and moods to shape their narratives. In contrast, P3 emphasized locations, events, narrative structures, and moods as key elements for expressing their authorial intent.

3.2 Challenges in Authoring Branched Narrative Fictions

Based on the formative interview, we summarized three design challenges **(C1-C3)** that participants faced when authoring branched narrative fictions.

C1: Writers block. A major challenge in iterating on branched narrative fictions is experiencing creative slowdowns when developing individual storylines, coming up with various branches, or devising diverse compelling endings. All participants mentioned this being a major hurdle when iterating on the branches. For example, P2 would have to rewrite the entire story branches or backtrack to overcome a creative rut and gain momentum.

C2: Difficulty in visualizing branched narratives. All participants expressed challenges in visualizing the story graph they were creating. Such a visualization would not only provide an overview of the branched narrative but also assist in brainstorming ideas and ensuring the overall viability of the storylines. P2, P3 manually map out the graph by sketching it on a paper (Figure 2), while P1 would write things out in a word document. Additionally, P3 would annotate individual nodes with notes to revisit later, aiding in the refinement and development of the storyline.

C3: Consistent verification. All participants noted a lack of effective verification tools for their creations, each with distinct priorities. For instance, P1 and P2 wanted to check for storyline consistency and confirm that branches were "correct" (e.g., matching locations), while P3 focused on ensuring that the story progressed meaningfully toward its climactic event. Despite these concerns, they generally relied on manual verification to spot potential flaws—especially after changing branches or adding events.

3.3 Design Goals

Based on the findings (Section 3) and the challenges **(C1-3)**, we established a set of design goals **(DG1-5)** to guide the development of WhatIF.

DG1 - Enable rapid creative experimentation to open up narrative possibilities. To overcome creative slowdowns, the system should provide alternatives and empower users to think creatively and expand branches **(C1)**. Additionally, the system should be equipped with ways to quickly open up narrative possibilities, such as transforming a linear storyline into branched storylines.

DG2 - Overview the branched narrative fiction. Developing branched narratives requires users to plan, track, and reflect [59] on the entire story graph. This helps with brainstorming, checking the integrity, and understanding the overall structure **(C2)**. The system needs to support visualizing and navigating the branching structure, such as zooming in and out to focus on specific segments, and searching for or retrieving specific nodes.

DG3 - Provide user-defined verification mechanisms. Manual verification after each change increases cognitive load and slows down their creative process **(C3)**. The interface should empower users to validate their creation using custom-defined parameters, ensuring their authorial intents aligns with the desired outcomes.

DG4 - Support iterative refinement. Branched narratives are created through the iterative expansion of one storyline to another. The system needs to support iterative refinement of the graph to allow for a continuous development based on their authorial intents **(C1)**. Users should be able to add, edit or delete event nodes **(C2)**. They should also be able to tag and write notes for future reference to build richer fiction.

DG5 - Showcase detailed view of each storyline. A root to leaf traversal of a BNF graph would lead to a single storyline. Users should be able to see the details of each storylines in terms of parameters commonly used as extracted from the formative study,

Figure 2: Artifacts from one of our formative study participant (reprinted with permission): (left) illustrates a branched structure used to develop branched narrative graphs, where the rectangles represent event nodes. (right) shows event nodes organized by acts, providing a clearer overview of the narrative's "big picture"

such as events, characters, locations and moods to aid in writing richer fiction **(C1, C2)**.

4 WhatIF: A Mixed-Initiative Tool for Authoring Branched Narrative Fiction

WhatIF is an interactive authoring tool (Figure 1) designed to help authors create, explore, and refine branched narrative fiction. By integrating a node-link visualization with large language model (LLM) assistance, it enables authors to iteratively develop and structure complex storylines while maintaining control over their creative intent.

Through an interactive graph-based interface, authors can modify story events, add branching paths, and adjust parameters such as mood, location, and character attributes, all of which dynamically update the narrative structure. Additionally, WhatIF provides storyline validation and thematic analysis, ensuring narrative coherence across branches.

Authors can start either from an empty canvas, manually adding and connecting event nodes to construct a linear story or by loading a written story, which the system automatically segments into event nodes using an LLM, placing them in the Authoring Panel for further refinement.

4.1 Authoring Panel: Graph-Based Narrative Development

Inspired by Twine [70], the Authoring panel (Figure 1A) provides an interactive node-link visualization of the narrative structure. Event nodes (Figures 1-a1 & 4) represent key moments in the story, with connecting links (1-a4) indicating branching paths. Links are color-coded based on the dominant mood of their source event. Link thickness encodes event likelihood, initially distributed equally among branches but adjustable via the + icon. The graph dynamically restructures itself as authors or the LLM-driven system make edits, ensuring a coherent visual overview of branching narratives. WhatIF actively supports narrative structuring, authoring, and iteration, providing ways to generate and organize branched narrative fiction.

4.1.1 Features for Narrative Generation (DG1,DG4). Event Nodes (Figures 4 & 1-a1) contain a high-level summary and editable description (Figures 1-a3 & 4), a list of characters, locations, and symbolism (Figure 1-a4), and functionality buttons (Figure 1-a2, 4)

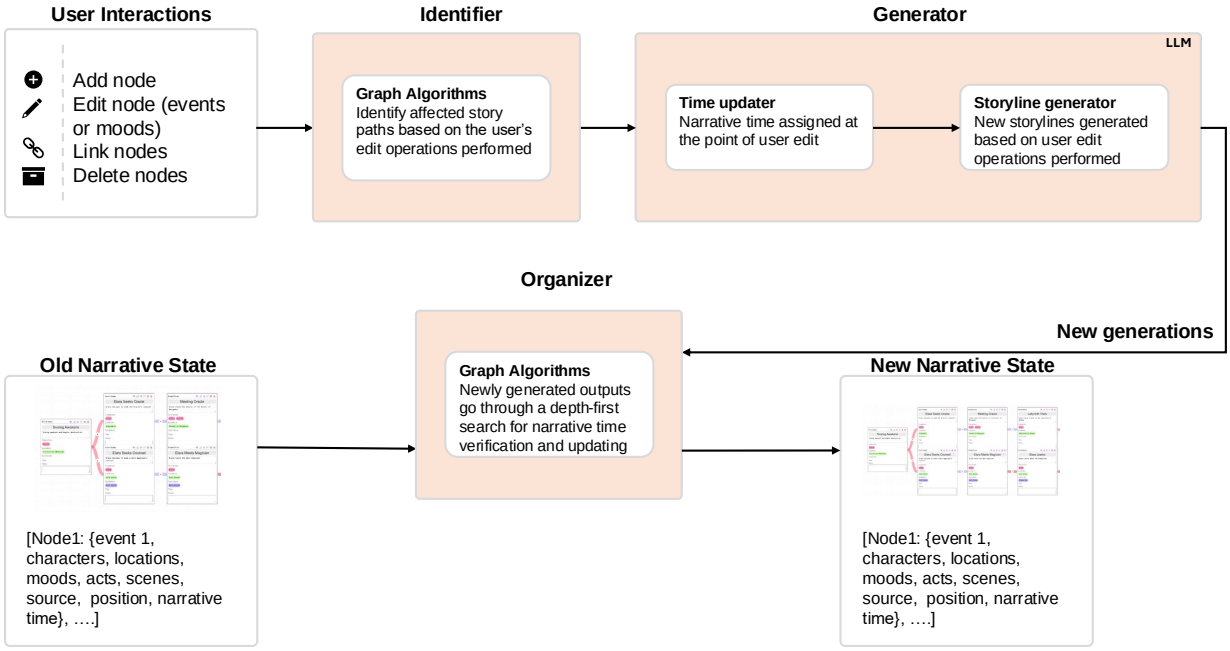

Figure 3: The workflow of WhatIF which involves three parts, an identifier which identifies the affected storyline based on user edits, a generator which takes in the user edit inputs and the previous narrative graph state to generate a new emergent storylines and an organizer which takes in the newly generated storyline along with the narrative time updates, applies depth first search to validate and update a new narrative graph state which is then displayed on WhatIF

to generate continue storylines, create new branches, edit moods, or create visual overviews over entire story paths. Modifications on a node propagate across all storylines where the node appears, reducing redundant edits and ensuring consistency.

Storyline Continuation. Authors can manually add, edit, or delete nodes. Additionally clicking "Play" prompts the system to generate three sequential events using the LLM, while ensuring that new content aligns with the existing storyline to this point.

Branch Recommendations. To aid brainstorming and creative expansion, clicking the "Fork" button generates three alternate story continuations using the LLM.

Both aforementioned features take the existing narrative structure—including surrounding nodes and connections—into account to maintain coherence.

Mood Blending. Authors can guide the tonal evolution of a storyline by adjusting eight foundational moods inspired from [1, 61] via a radar chart with sliders (Figure 5). Clicking "Play" applies the updated mood profile to newly generated events, allowing for thematic experimentation.

4.1.2 Features for Narrative Organization (DG2,DG5). To maintain clarity in complex branching structures, WhatIF automatically organizes event nodes on the graph, with the x-axis representing narrative time [30] ensuring chronological coherence, and the y-axis sorting by edit recency, keeping the most recently modified path at the bottom.

Path Explorer. As the graph grows, visual clutter might obscure storylines. Clicking the "Path" icon on an event node extracts all paths containing that node, displaying them in the Storyline Panel (Figure 1-C) for easy exploration (see section 4.2).

Node Tagging. To enhance organization, event nodes can be labelled with color-coded tags, allowing authors to highlight themes, track character arcs, or customize it based on their needs.

4.2 Storyline Panel: Thematic and Structural Analysis.

The Storyline Panel (Figure 1B) provides a deeper view of story branches selected in the Authoring Panel, featuring (1) a Mood River, visualizing thematic evolution, and (2) storyline bars for quantitative story analysis.

Mood River visualized thematic moods of the narrative graph selected in the Authoring Panel. The x-axis represents narrative time [30] while the y-axis displays storyline branches containing the chosen node. Events appear as color-coded rectangles (Figure 1-b1)–orange for user-authored, blue for LLM-generated–with shared events merging into a single block to highlight common plot points.

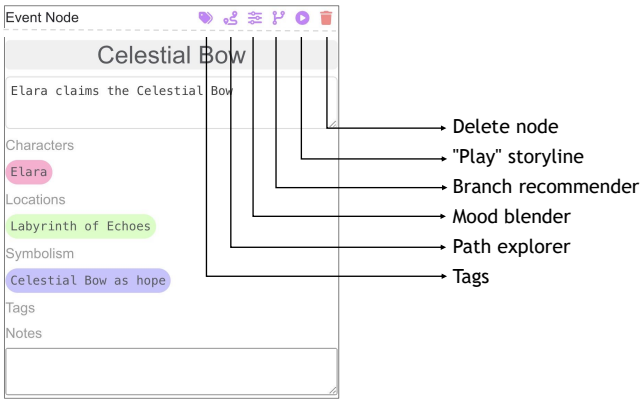

Figure 4: Storylines are made up of Event Nodes, with each node consisting of the event's title, a textbox for event description, as well as the characters, locations, and symbolism associated with the event.The event node's main main functionalities are located at the top of the node, including "Node Tagging", "Path selector", "Mood blender", "Branch recommender", "Play node", and "Delete node". A text box labeled "Notes" is present at the bottom.

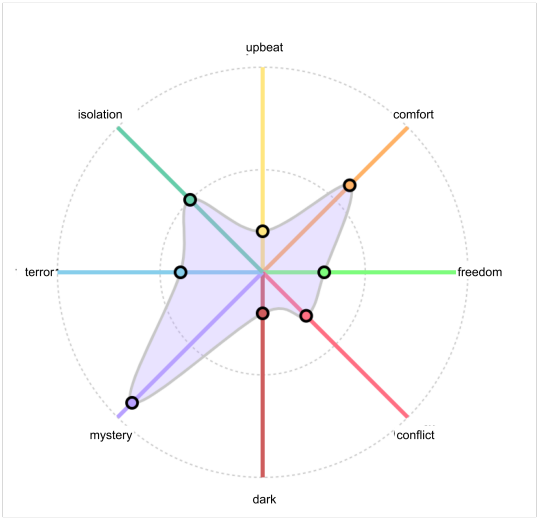

Figure 5: The mood blender radar chart visualization which provides axial sliding to adjust and "blend" different moods. The chart visualization has eight labelled dimensions: upbeat, comfort, freedom, conflict, dark, mystery, terror, and isolation. Each axis has a circular node that can be adjusted along its length.

Links between rectangles encode eight mood dimensions (Figure 1-b2), with thickness indicating intensity. Hovering over elements highlights the corresponding event in the Authoring Panel, displaying additional details in a tooltip. This view enables authors compare tonal shifts, iteratively refine moods, and validate how well storylines align with their intended emotional arcs.

Storyline Bars (Figure 1-b3) provide a numerical overview of each selected storyline, aligned with the Mood River's vertical axis (Figure 1-b3). The x-axis (ranging from 0 to 1) reflects different metrics based on user focus. By default, it identifies each storylines dominant thematic mood to determine primary genre. If storyline verification is enabled (explained shortly), the x-axis represents user-defined metrics, allowing authors to assess thematic consistency while referencing mood trajectories.

4.3 Location and Character Panel: Visualizing Character Interactions across Locations. (DG5)

Inspired by prior work on storyline visualization[8, 30, 67, 75], this panel helps authors track character interactions across locations over narrative time (1). Users compare two storylines side-by-side, see when and where characters interact with overlapping events highlighted, and identify disruptions or inconsistencies in character movement.

4.4 Storyline Verifier: Evaluating Narrative based on User-defined Metrics (DG3).

To help validate narrative structures, users can apply custom metrics (e.g., coherence, consistency) with their own definitions, which uses another LLM as a judge to score each storyline on [0–1] scale. Scores are visualized as bar chart on the Storyline Panel (Figures 6 & 1-b3), providing authors with insights into how well the storyline aligns with their metrics, allowing for targeted iterations.

4.5 Implementation Details

Our prototype follows a client-server architecture. The frontend, built with D3.js, React and JavaScript and uses ReactFlow to support the interactive graph layouts in the Authoring Panel. The Flask server handles all user interactions and backend requests. The backend is built using Python and Flask, using GPT4-turbo via the OpenAI API.

Narrative updates are generated through a three-stage process tightly integrates LLM-based generation with graph based narrative structuring. When a user modifies a story node–by editing events, adjusting moods, deleting event nodes, altering graph connections, or possibly a combination of these actions–the system aggregates these inputs along with the initial seed story (if present), the current graph structure, the spatial location of the modified node, and the narrative path leading to the change. This graphical information is first passed on into a DFS which initially performs the task of identifying the relevant storylines that are affected by the user edits performed. The entire contextual information along with these affected storylines is then passed into the LLM which, based on the user actions, provides an updated narrative and specifications for the newly generated story nodes (e.g., characters, locations, symbolism). To maintain coherence and fix inconsistencies that might potentially be present in the narrative time after the LLM generation, the output is again processed using Depth First Search (DFS) algorithm, dynamically reconstructing the graph layout. The x-axis reflects narrative time [30], ensuring chronological consistency, while the y-axis prioritizes recently edited nodes, providing

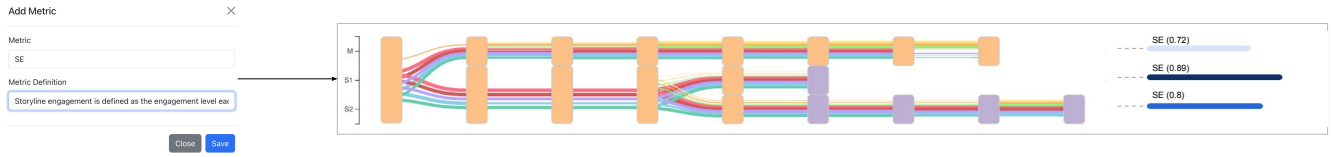

Figure 6: Storyline Verifier lets users name and define a custom metric which showcases metric scores for each selected storyline in the Storyline panel. The left side of the image shows the modal with input fields labelled "Metric" and "Metric Definition," where "SE" (Storyline Engagement) is being defined. Below the text input fields, there are "Close" and "Save" buttons. An arrow points from the modal to the right side, which displays a Sankey diagram with multiple colored flows connecting rectangular nodes, representing different storylines. To the right of the diagram, three horizontal bar charts display "SE" scores (0.72, 0.89, and 0.8) for different selected storylines, with varying shades of blue indicating different engagement levels.

an intuitive spatial representation of updates in the authoring panel. See Figure 3 for more details.

Furthermore, to aid in custom storyline verification in form of the storyline verifier, we use another instance of the same LLM model with a new conversation history as a judge model inspired from prior works [39, 60]. Since the creative process of crafting interactive fiction is highly dynamic and subjective, traditional metrics like BLEU[51] or ROUGE[37] scores fail to capture nuanced qualities such as coherence, character arc development, consistency or any other metric a human would consider in their creative process. Given their training on vast human-authored corpora, LLMs can serve as effective proxies for human judgment, motivating our use of an LLM as a judge model. This LLM is provided with the selected storylines, the metric, and its definition for measurement and is tasked with evaluating the storylines based on the given criteria, assigning a rating between 0 and 1, where 0 represents the lowest score and 1 represents the highest.

5 Use Case

To demonstrate how WhatIF supports the creation and exploring branched narrative fiction (BNF), we present a scenario featuring Gary, an indie game developer wanting to outline plots for his next game. Figure 7 illustrates his workflow. The scenario is inspired by patterns observed during our user study.

Gary begins by loading a seed story with a *Good vs Evil*, which WhatIF automatically breaks into its event nodes in the Authoring Panel. The story follows Elara, a hero who must defeat Svarog, a dragon that threatens her village, Greendale. Guided by an Oracle, Elara retrieves a celestial bow from the Labyrinth of Echoes and ultimately defeats Svarog in a climatic battle.

5.1 Expanding the Narrative Graph

Seeking greater creative variety, Gary expands the storyline by introducing an alternate path: Instead of meeting the Orcale, Elara encounters a dark magician in the Realm of Darkness. Clicking Play he prompts the system to generate three new event nodes continuing this storyline.

However, the new branch–*Elara finds a Shadow Arrow and defeats Svarog*–mirrors the original plot, offering little diversity. Stuck in a creative rut, Gary deletes the repetitive nodes and uses the Fork tool to request alternative branches.

This generates three new storypaths: *Experimenting with the arrow's power*, *Using the arrow against Svarog* mirroring the original outcome), and *Elara discovers the Shadow's Curse*. Intrigued by the last option, he clicks that node's *Play* button to continue the story. This leads to a bittersweet ending where Elara deveats Svarog but falls into an eternal slumber.

Wanting to explore a more tragic outcome, Gary edits *Using the Arrow against Svarog*, modifying it to *Elara fights Svarog, but things take a turn for the worse*. Upon clicking Play, the system continues the story and generates events in which Elara dies and Svarog takes control of Greendale.

5.2 Iterating on Thematic Moods

To ensure emotional variety across branches, Gary uses the Path Explorer tool to populate the Mood River, visualizing how emotional tones evolve over narrative time. He notices that the new paths–where Elara dies or falls into an eternal slumber–are dominated by dark, isolation, and conflict-heavy moods, while the original storyline maintains a more mysterious tone. The Storyline Bars further highlight these differences.

Seeking more diverse emotional arcs, Gary selects the tragic storyline and uses the Mood Blender to adjust its composition. Lowering terror, isolation, and darkness, while increasing comfort and upbeat, he shifts the narrative's tone.

After applying the changes, Gary he observes a hopeful ending: Elara's sacrifice transforms Svarog, leading him to repent and rebuild Greendale. The Mood River updates, showing a balanced emotional trajectory across branches. By iterating on mood composition, Gary ensures that his branching narrative offer distinct, yet cohesive, emotional experiences.

5.3 Addressing an Anticlimactic Storyline

Curious about character interactions, Gary uses the Location-Character panel to compare two endings: one where Svarog rebuilds Greendale and another where Elara enters eternal slumber. He notices that in the former Elara and Svarog's interaction is brief, making it feel anticlimactic. To improve pacing, he inserts a new event where Svarog's minions attack Elara, creating a battle sequence before facing Svarog. This revision results in a more dynamic and satisfying conclusion.

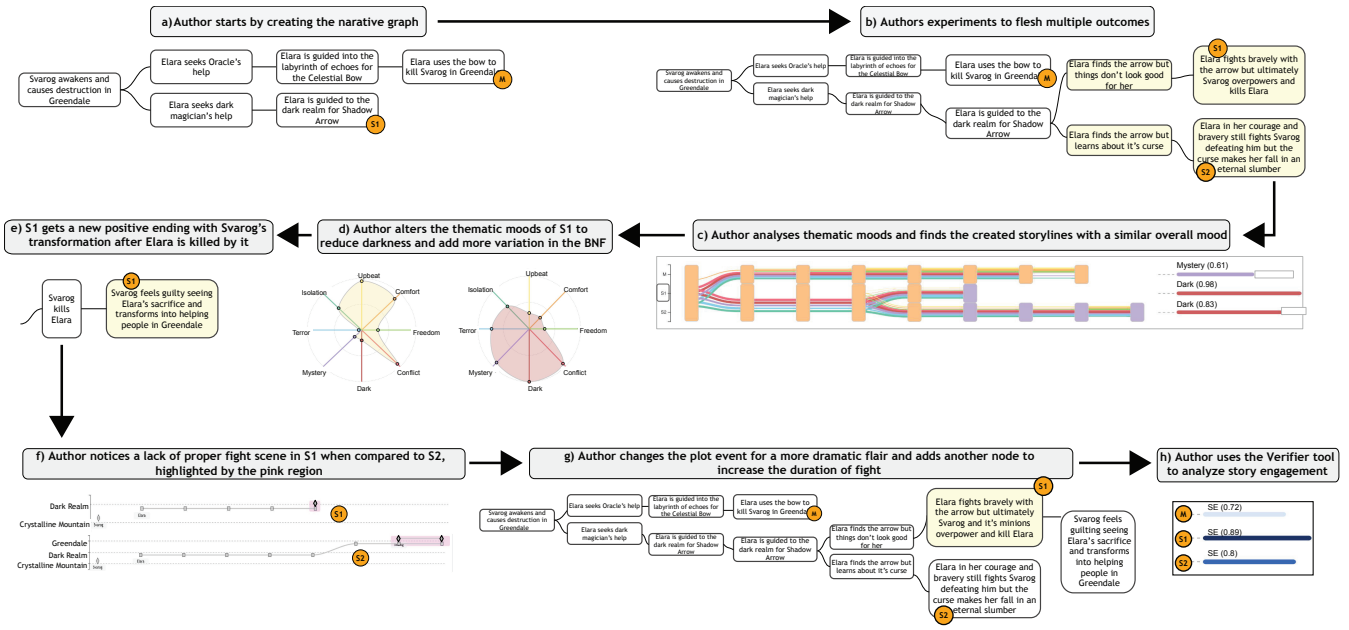

Figure 7: An example workflow that shows a user creating 2 new storylines (S1, S2) inspired from the main/seed storyline (M). (a) The user starts out with manually adding an alternate event, (b) creating 2 new branches, thus 2 new storylines by leveraging the branch recommendations feature provided, (c) analysing the thematic moods for all the 3 storylines to see for thematic diversity, (d) altering moods for (S1) to a more positive ending as the story is too dark and S2 showcases similar themes across it's storyline, (e) leveraging GenAI to flesh S1 given altered moods, (f) analysing locations and character for the newly generated storylines (S1, S2) and noticing a lack of good conflict in S1, (g) further manually adding a node to increase conflict between the main characters and finally, (h) using custom metric of *Storyline Engagement* to verify all the 3 storylines (M, S1, S2). Newly generated nodes are highlighted in yellow. Note: the UI elements in this figure have been simplified for clarity, for true representation of event nodes please see Figure 1

5.4 Evaluating Storyline Engagement

With the BNF graph finalized, Gary wants to assess audience engagement using the Storyline Verifier tool. He defines a metric, *Storyline Engagement*, as *"Audience is emotionally invested in a cohesive narrative"*. The Mood Panel updates to display engagement bars, with values ranging from 0.72-0.89 also shown as colors. The highest scoring branch–the bittersweet*eternal slumber* ending–suggests its emotional complexity and plot twists enhance engagement. Satisfied with these insights, Gary finalizes his branched narrative and prepares to develop his next indie game.

6 Evaluation

We conducted a user study with 11 participants to assess how WhatIF supports users in crafting, experimenting, and analyzing branched narratives as well as the challenges users faced with the LLM-assisted tool. During the study participants were asked to author a BNF using *WhatIF* first in a controlled task and then through freeform exploration. We explain the study protocol in more detail below.

6.1 Branched Narrative Creation Task

Through a controlled task users were asked to author a BNF, including story creation, editing, and verifications, while following the think-aloud protocol. A structured task ensured engagement with core system features, which might be overlooked in freeform exploration. Participants started with a preloaded *"Good vs Evil"* story featuring a hero and a dragon (Section 5). Participants were guided to: 1) manipulate the BNF graph (adding 2-3 nodes, linking nodes, and deleting nodes); 2) use the LLM-based tool to play out storylines and obtain branch recommendations; 3) analyze the BNF using Mood River and Location-Character Overview panels; 4) verify the BNF using metrics of their choosing.

6.2 Freeform Exploration

During this part, participants built out their BNF from a set of initial storylines, engaging in open-ended exploration while encouraging meaningful system interaction. Their objective was to design a BNF with 3-4 distinct endings, evenly balancing "good" and "bad" outcomes, based on their own interpretations. Participants were encouraged to think out loud and freely interacted with all available panels and features.

6.3 Participants and Procedure

We recruited 11 participants via an internal Slack campaign and through IntFiction.org, a forum for interactive fiction. Participants had an average of 7.5 years experience in BNF development either as Dungeons and Dragons (DnD) Masters or as hobbyist/professional

narrative fiction creator (Table 3). The study was conducted via Zoom, lasted 90 minutes per session, and participants received $100 USD as compensation. The study protocol was approved by our institutional ethics review board.

Pre-Study Interview. A semi structured interview explored participants' motivations, design principles, prior experiences, and challenges in BNF development.

Tutorial. The study administrator provided an overview of WhatIF, guiding participants through BNF creation, visualization, verification, and queries. Users controlled the system via Zoom screenshare with guidance from the study administrator.

BNF Creation Task. Participants completed a controlled BNF creation task (section 6.1), thinking aloud to share their thought processes during each step .

Freeform Exploration. Participants freely explored the interface, using either provided prompts or their own ideas, with their interactions recorded and while following the think-aloud protocol.

Post-Study Questionnaire and Interview. Upon completing the freeform exploration, participants rated system features using a Likert scales. We administered an unweighted Creativity Support Index (CSI) [15] to assess the perceived system's effectiveness in fostering creativity. A semi-structured interviews gathered qualitative feedback on overall experience, interface design, and challenges they encountered.

6.4 Quantitative Findings

We report the self perceived usability scores of our WhatIF's core functionalities and also an unweighted Creativity Support Index (CSI). Overall, we found that (1) the core functionalities (graph layout, branch editing, and AI-generated suggestions) were rated highly, indicating that users found the system intuitive and easy to use for structuring interactive narratives. (2) Users felt creatively supported, especially in terms of enjoyment and the value of their outcomes. The system effectively aided exploration and expressiveness during short-term use.

6.4.1 Feature Usability Ratings. Participants overwhelmingly rated the core functionalities of the WhatIF interface as useful. In particular, the authoring panel, narrative generation capabilities, and editing tools all received high ratings (M≥4.00), indicating that users found these features both intuitive and highly effective for generating and organizing narrative fiction. World-level visualizations received mixed feedback. The location-based view (M = 3.89, SD = 0.57) was rated more positively, while the mood-based view (M = 3.67, SD = 0.94) scored lower—possibly because users didn't have enough time to fully explore it during the session. The verifier tool (M = 3.56, SD = 0.83) also got moderate ratings. These findings indicate world-level visualizations may require more fine-tuning for studying user preferences in longer-term usage.

6.4.2 Creativity Support Index (CSI). To assess broader creativity-related outcomes, we adapted the Creativity Support Index by using unweighted Likert-scale ratings, omitting the Collaboration factor in line with prior work [15, 63, 64]. This was also adapted to reduce participant burden in the user study. This unweighted CSI thus serves as a simplified measure of perceived creativity support.

Table 1: Mean and Standard Deviation of perceived interface ratings. Most questions correspond to specific design goals from Section 3.3. The highest value is in bold. The raw scores are showcased in Figure 8.

Question	Mean (*M*)	Standard Deviation (*SD*)
Q1: Visualization of the graph layout was useful (DG2)	**4.56**	0.50
Q2: Adding, deleting, and linking branches were useful (DG4)	4.22	0.92
Q3: The AI's branch recommendations helped me think in newer ways (DG1)	4.00	0.67
Q4: Seeing how the worlds differed in terms of moods (DG5)	3.67	0.94
Q5: Seeing the number of nodes generated by me vs the AI	3.44	0.96
Q6: Seeing how the worlds differ in terms of locations (DG5)	3.89	0.57
Q7: The verifier helped me see how well I was performing (DG3)	3.56	0.83

Table 2: Unweighted Creativity Support Index (CSI) Results. The highest value is in bold. Since our study did not involve collaboration, we followed a similar method from [15, 63, 64] and omitted the Collaboration Factor. The raw scores are showcased in Figure 9.

Factor	Mean (*M*)	Standard Deviation (*SD*)
Exploration	3.78	1.18
Expressiveness	3.94	1.15
Immersion	2.89	1.15
Enjoyment	4.11	0.99
Results Worth Effort	**4.22**	0.79

Moreover, our results underscore that the WhatIF interface provides strong creative support. As shown in Table 2, users rated the interface favourably in terms of enjoyment (M = 4.11, SD = 0.99) and felt that the outcomes were well worth their efforts (M = 4.22, SD = 0.79). In contrast, the Immersion factor received a notably lower rating (M = 2.89, SD = 1.15), suggesting that users remained consciously aware of operating a LLM-assisted tool rather than becoming fully absorbed in the creative process.

6.5 Qualitative Findings

We next report comments and feedback collected throughout the study and after the freeform stage. We performed an open coding on participant verbalizations (think aloud and additional commentary), and discuss both positive feedback as well as some suggested system improvements below, in the context of the WhatIF's design goals. We also report the perceived interface ratings when needed.

6.5.1 WhatIF supported rapid narrative experimentation (DG1). Many participants praised WhatIF for its ability to rapid flesh out

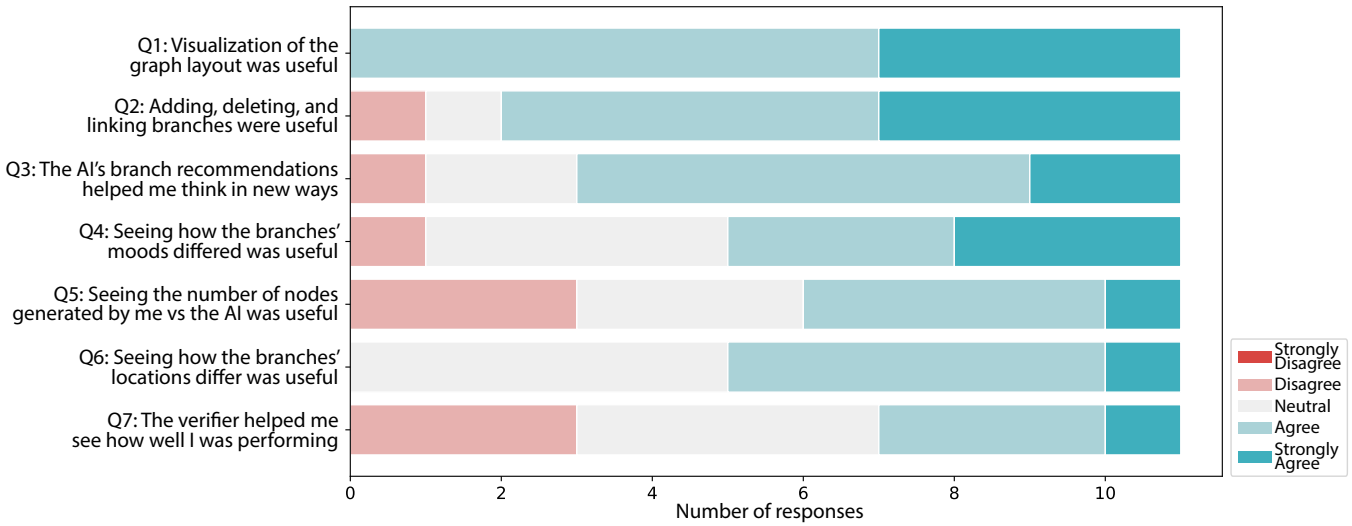

Figure 8: Post study survey results from our study participants, rating the usefulness of WhatIF Perceived interface ratings (N=11)

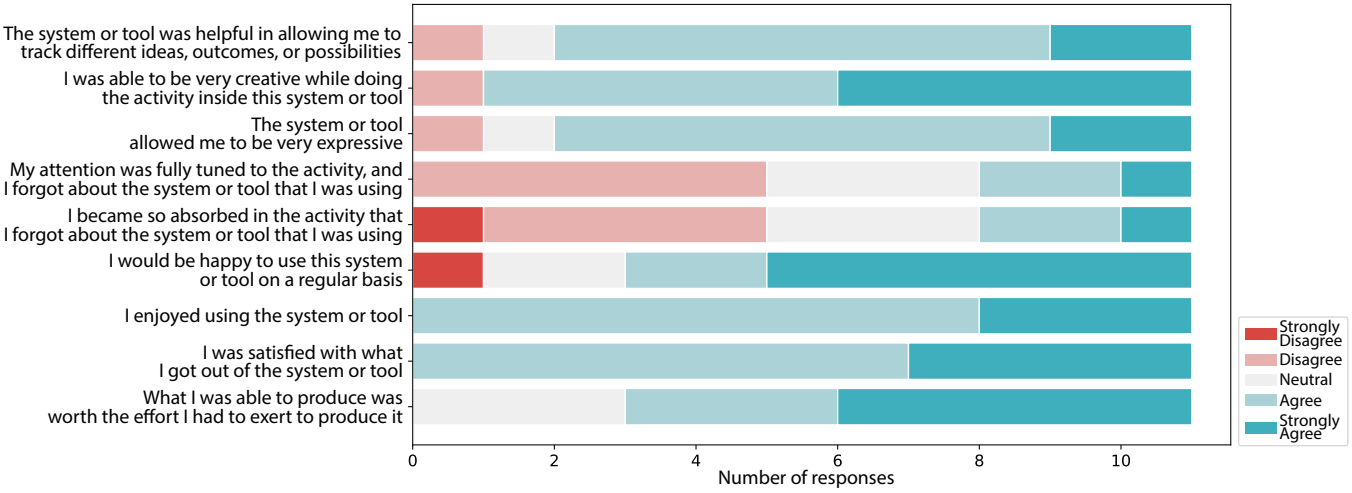

Figure 9: Participants' post-study responses to the Creativity Support Index, assessing perceived effectives in fostering creativity (N=11)

storylines based on their intents (P11, P8, P7, P6, P4). They also commented the branch recommendation to be highly useful for experimenting with various ideas and helping them get rid of writers block. This sentiment was frequently expressed by participants when editing nodes, creating branches, or obtaining branch recommendations. This is also reflected in the high score for AI's help to think in new ways (Table 1-Q3, M=4.00,SD=0.67). For instance, P11 noted: *"This is an amazing planning tool, it helps me check and see many storylines quickly"*. This highlights WhatIF's strength in quickly generating multiple possible alternatives and help users imagine what might be possible.

6.5.2 Authoring panel provided useful temporal node-link layout visualization (DG2). All participants who participated in the study had mentioned a lack of tools to overview the creation. Three participants who have published interactive fictions have used existing tools such as Twine [70] and Inform7 [23], while other participants either sketched out the fiction using pen-paper or excel sheets. However almost all the participants found the authoring panel to be the most useful for visualizing events and their connections. This was mainly because WhatIF relieved the users of the burden of arranging storylines temporally, which aided in making coherent sense over existing tools like Twine that lack such functionality. *"It is empowering to see all the story nodes and links laid out. It's sort of helping me think of ways I want to further explore this tool"*(P9). *"I can put out broad strokes of the story and see how it emerges, that's really interesting "*(P6). This positive sentiment is also reflected in the perceived interface rating of the authoring panel (Table 1-Q1 -

M=4.56, SD=0.5). A potential challenge for this temporal node-link layout would be visual cluttering as the graph grows: *"I really like the visualization, however I wonder how big will it get if I put in Game of Thrones"*(P7). While users could zoom in and out to view their creation, the increasing complexity of the visualization made it difficult to maintain clarity and usability for more intricate narratives. Two participants (P1,P5) suggested a resizeable authoring panel to provide a larger canvas for planning their storylines, which we discuss further in Section 7.2.

6.5.3 Mood river enhances narrative control but requires fine-tuning for user preferences (DG4, DG5). Participants highlighted the significance of mood visualization in maintaining the thematic tone of their narratives. Five participants emphasized using mood visualization to ensure the overall narrative aligns with their creative intents (P11, P10, P9, P7, P5). *"I definitely consider moods to set the tone of a DnD session, so if I see a specific mood getting heavy I would use the visualization to possibly add other moods depending on the overall narrative"*(P11) *"I find the visualization really useful as I would use this information of moods for various storylines to guide the story progression and possibly tailor it for different player groups when I am DMing"*(P9). Others valued the mood visualization to aid them in developing characters in a narrative and for maintaining thematic consistency. Participants (P10, P11) found the mood panel allowed them to visualize and reinforce the desired themes, such as creating a horror atmosphere by focusing on specific moods.

Interestingly, while participants consider mood important, most do not call it out explicitly in their process. As one participant noted, *"I do consider moods when I am a dungeon master, however it's kind of implicit, I have never explicitly classified moods into categories"* (P11). Another participant echoed a similar sentiment, *"This visualization is very interesting, I have never classified moods when I create an IF, but this is interesting to see"* (P7). However, some participants found the mood panel visualization to be overwhelming (P11, P5, P1). *"The visualization is a bit overwhelming as I am seeing it for the first time"*(P11). Two participants wished for finer control (P5, P4). *"I would have appreciated a more fine grained control over moods rather than just changing slider values, possibly by prompting"* (P4). This is reflected in a relatively lower average (M=3.67) and a higher standard deviation (SD=0.94), see Table 1-Q4.

6.5.4 Location-Character panel helped users progress the storylines (DG4, DG5). Participants widely recognized the location character panel as a pivotal feature for organizing character placements and interactions within their narratives (P1, P4-5, P7-11), also reflected in the score (Table 1-Q6, M=3.89, SD=0.57). Participants appreciated the clear and and concise visualization, which offered a structured representation of characters and location. *"Personally, for me the pivots in storylines are determined by locations and characters, so seeing this laid out like this is really useful"*(P7). This helped to maintain narrative consistency and facilitate dynamic interactions, compared to their prior experience, which required manual tracking. *"This is really useful, I would use this to determining what character interactions to make for the storylines"* (P10). Participants also used the location character panel to help them alter the authoring panel by either adding more character interactions or progressing plot point given the presence of characters in specific locations - *"I can now clearly see how many characters interact where, so based on the location I will now be able to add a new side plot at this timestep"*(P7). *"See at this timestep there's only one character, I think I would add additional characters here to make the plot interesting"*(P5). Participants valued the panel for enhancing their world-building efforts and ensuring narrative coherence, *"Useful for maintaining coherence in storylines, if everyone is scattered across locations can make sure to not add faulty interactions"*(P4). Participants (P11, P9, P6) suggested additional features to support their workflows, such as storyline visualization from each character's point of view, or LLM-assisted character design.

6.5.5 Storyline verifier enables automated assessment of branched narratives (DG3). While most participants had never used such a feature (P4-5, P7-11), they appreciated storyline verifier for providing automated assessments that complemented their manual processes. Interestingly, participants came up with different metrics, such as hilarity, predictability, and character arc development, each with their unique definition. *"Generally there's a lack of such tools, but this is an added benefit to check different ratings"*(P11). P4 highlighted its utility as an assistant - *"I would definitely use the verifier, however I would still perform manual checks. And if the LLM scores something as low, I would go back and then verify again"*. Participants valued the storyline verifier for saving time and offering objective evaluations. However, three participants (P6, P9, P10) expressed reservations about the tool's necessity, as mentioned by P6 - *"Don't really think this is useful, because if writing professionally one should know if story is good or not."*. The novelty of such a feature might challenge users on incorporating it in their unique workflows - *"Useful feature but need time to adapt to this"*(P9).

7 Discussion

Our study demonstrates that WhatIF enhances branched narrative creation. Participants highlighted that the visual authoring panel provided an intuitive, high-level abstraction of story structures, enabling rapid prototyping and iterations. In addition, the branch recommendations expanded narratives possibilities with inspirations. Other panels–such as mood river and location-character overviews–allowed users to contextualize thematic tones and spatial relationships within their narratives. These multi-faceted features yielded high ratings on measures of expressiveness and results from the unweighted Creativity Support Index (Table 2). In this section, we discuss the implications from user studies, as well as the limitations and potential directions for future work.

7.1 Mixed perception of control with AI

Participants were aware they were interacting with an AI-driven system. This awareness was heightened because features within WhatIF relied on API calls LLM, which occasionally resulted in delays, contributing to a lower immersion score(M=2.89, SD=1.15). Despite these interruptions, participants generally exhibited understanding when the interface failed to produce the correct output. A key concern was the limited control over LLM outputs. Five participants (P1, P6, P7, P10, P11) noted that the generated storylines often lacked diversity because the system constrained the LLM to closely follow the initial narrative to avoid excessive branching. Thus expressing a desire for more control over this "super prompt" suggested additional features like a character profile database and

location-based suggestions—insights that will inform future iterations.

Interestingly, despite these concerns, all the participants felt they were able to express their intents well, over created storylines highlighted by a relatively high Expressiveness score (M=3.94,SD=1.15) (Table 2) in which LLMs assisted them when asked at the end of their study. This dual perception suggests that while participants felt a need for more advanced features to tune their narratives, they also acknowledged the current tool's effectiveness in supporting their creative processes. This highlights the need for more nuanced mechanisms in WhatIF which can keep the LLM control "abstract" but also provide an option to fine tune the prompts based on users needs.

7.2 Hierarchical graph representation for better clutter management

Our evaluations (Section 6) show that while the authoring panel narrative generation was found to be highly effective, interface clutter emerged with more than four storylines (7–8 nodes each). Although mechanisms such as zooming, panning, and clustering by acts or chapters were available, they were underutilized, likely due to limited exploration time. In future iterations, we plan to make these controls more prominent and integrate a hierarchical graph visualization that adaptively clusters nodes based on the user's zoom level, inspired by [22, 44].

7.3 Human intervention needed to add tension in storylines.

A recurring theme was the lack of tension in LLM-generated storylines. Even when participants shifted to darker themes like conflict or terror, the model maintained these only for 2–3 event nodes before reverting to optimistic outcomes. P4, P10 observed similar behavior in ChatGPT, while P5 achieved better results by explicitly prompting (e.g., "do not do this or do that") to sustain a darker thematic direction. Some users increased the number of "undesirable" events to force a darker narrative, echoing prior findings [12, 68] that human intervention is often necessary to generate tension-filled storylines. In future work, we aim to develop interaction techniques that consistently guide models to produce more tension and conflict.

7.4 Balancing Structure and Creative Flexibility in BNF Design

Our formative interviews showed that creative processes are rarely streamlined, highly iterative and often cumbersome. Some participants began with a high-level pivot story point, while others started with abstract concepts like mood or overarching themes. This divergence sometimes elicited polarizing feedback in user studies; one participant remarked, *"I am enjoying this tool so far. Now that I have seen this, it is hard to go back to the manual process I do"* (P10), a sentiment echoed by P2 and P5, whereas P9 felt constrained by the interface's methodical approach.

Most participants appreciated WhatIF's features but also suggested enhancements tailored to their workflows—P12 wished to view storylines from each character's perspective, and P14 called for an LLM-powered character database to enrich character development. These contrasting views highlight the tension between offering supportive structure and maintaining creative flexibility. A promising solution is a personalized interface that adapts to individual creative processes, leveraging generative AI to rearrange components (e.g., character motivations, thematic arcs, mood boards) without constraining creativity.

7.5 WhatIF Leaning into a Narrative Planning Tool.

Many participants found WhatIF useful for structuring storylines, with some suggesting adding Mural-like whiteboard functionality. They envisioned AI-assisted transformation of loosely structured ideas into cohesive, branched narratives, easing cognitive load and enhancing creative exploration.

Although this could enhance creative flexibility, it also introduces AI-driven planning challenges. Our experiments with GPT-4 for automated event generation between defined start and endpoints revealed issues with coherence and contextual consistency. Future iterations may benefit from reinforcement learning–enhanced models (e.g., OpenAI's O1, DeepSeek's R1) to improve narrative structuring for BNF authors.

7.6 Limitations in evaluating WhatIF

The small sample size in both the formative and user studies provides preliminary insights into the challenges of integrating AI into creative writing, as well as identifying features that may enhance or hinder the user experience. While 13 participants were initially recruited, two data points were excluded due to observed suspicious behavior. Expanding the participant pool, particularly with individuals experienced in writing non-linear narratives, would offer richer, more diverse insights into how WhatIF's features facilitate the creation of cohesive storylines.

Given the exploratory nature of our study, we focused on capturing first-use insights into the potential benefits and challenges that users face when interacting with WhatIF. This led us to adopt a qualitative approach, as opposed to an objective, quantitative one, due to the inherently subjective nature of creative works. Defining "correctness" in creative writing is difficult, as users' intentions in constructing storylines are shaped by personal experiences and are not easily evaluated against objective ground truths.

To gain more "objective" measures of success, future studies could include comparative evaluations with existing non-linear storytelling tools such as Inform7, Twine or ChoiceScript. However, challenges exist in this approach: many non-linear storytellers may not be familiar with coding or graph-based interfaces, and the lack of AI support in these tools would make them an unfair comparison to WhatIF. Nevertheless, we plan to conduct future comparative studies to better understand user behavior in both WhatIF and other AI-assisted tools for creating branching narrative fiction.

7.7 Ethical considerations of using GenAI for fiction development

In developing WhatIF, we acknowledge the ethical challenges of using generative AI in storytelling. Our goal is to augment—not

replace—human creativity by assisting in narrative generation and organization. While AI can craft compelling narratives, it often produces homogenized, overly positive outputs due to its lack of lived experience, emotional depth, and cultural context [68]. Additionally, inherent training data biases may influence the results. To mitigate these issues, WhatIF enables users to override or refine AI suggestions and incorporates providing a measure of control over the creative process.

8 Limitations and Future Work

While WhatIF demonstrates considerable promise in augmenting the creative process, our findings indicate a few challenges that merit further discussion — challenges that are trade-offs inherent in our design choices. Few participants noted that the AI occasionally produced storylines with limited diversity. This approach, while sometimes constraining creative divergence, was a deliberate decision in the design process to ensure that the narrative structure remained manageable and aligned with established storytelling conventions. Future work can consider combining symbolic planning with LLM to enforce narrative structure while maintaining content diversity. Additionally, WhatIF's authoring panel became cluttered as their narratives grew in complexity, which calls for more sophisticated ways to not only visualize large graphical structures but also manage the information present in it, such as LLM-powered zooming and summarization. Lastly, while the LLM-based verification tool was not universally embraced as a component of the creative workflow, it was intentionally positioned as an auxiliary resource to provide objective, supplementary feedback rather than replacing human judgement. Future work can examine ways to embed in author workflow. Overall, these challenges underscore the difficult balance between control and automation in AI-assisted creative systems.

9 Conclusion

Generative AI enables real-time content generation for branched narrative fiction, but tools for structuring complex storylines and controlling AI outputs remain limited. WhatIF integrates AI with interactive visualizations to augment narrative fiction authoring, allowing authors to transform linear stories into rich, branches narratives through event-based prompts and a dynamic graph interface, while keeping authorial intent central. A user study with 11 narrative developers demonstrated WhatIF's effectiveness, through mood-based and location-character visualizations, and LLM-assisted metrics verification, which helped maintain coherence and foster creative experimentation. These findings underscore the value of a mixed-initiative, human-in-the-loop approach, bridging creative ideation with technical execution and advancing branched narrative authoring.

References

[1] [n. d.]. Mood and Tone in Literature. https://www.cbsd.org/cms/lib010/PA01916442/Centricity/Domain/1793/mood-and-tone.pdf. Accessed: 2025-04-14.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[3] Choose Your Own Adventure. 2025. Choose Your Own Adventure. https://en.wikipedia.org/wiki/Choose_Your_Own_Adventure. Accessed: 2025-02-06.

[4] Wolfgang Aigner, Silvia Miksch, Bettina Thurnher, and Stefan Biffl. 2005. PlanningLines: novel glyphs for representing temporal uncertainties and their evaluation. In *Ninth International Conference on Information Visualisation (IV'05)*. 457–463. doi:10.1109/IV.2005.97

[5] Alberto Alvarez, Jose Font, and Julian Togelius. 2022. Story designer: towards a mixed-initiative tool to create narrative structures. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*. 1–9.

[6] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7375–7382.

[7] Paul André, Max L Wilson, Alistair Russell, Daniel A Smith, Alisdair Owens, and MC Schraefel. 2007. Continuum: designing timelines for hierarchies, relationships and scale. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. 101–110.

[8] Dustin Arendt and Meg Pirrung. 2017. The "y" of it Matters, Even for Storyline Visualization. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 81–91. doi:10.1109/VAST.2017.8585487

[9] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[10] Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints. *arXiv preprint arXiv:2410.04197* (2024).

[11] Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324* (2021).

[12] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–34.

[13] Charisma.ai. 2025. Charisma.ai. https://charisma.ai/. Accessed: 2025-01-30.

[14] Qing Chen, Shixiong Cao, Jiazhe Wang, and Nan Cao. 2023. How does automation shape the process of narrative visualization: A survey of tools. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[15] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[16] Frederik Roland Christiansen, Linus Nørgaard Hollensberg, Niko Bach Jensen, Kristian Julsgaard, Kristian Nyborg Jespersen, and Ivan Nikolov. 2024. Exploring presence in interactions with llm-driven npcs: A comparative study of speech recognition and dialogue options. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[17] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[18] John Joon Young Chung and Max Kreminski. 2024. Patchview: LLM-Powered Worldbuilding with Generative Dust and Magnet Visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[19] Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[20] Maryna Mykolaivna Dunaieva and Iryna Anatoliivna Morozova. 2016. The Peculiarities of Creative Writing, Its Characteristics, Typical Difficulties and the Way to Overcome Them. (2016).

[21] AI Dungeon. 2025. AI Dungeon. https://aidungeon.com/. Accessed: 2025-01-31.

[22] Niklas Elmqvist and Jean-Daniel Fekete. 2009. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE transactions on visualization and computer graphics* 16, 3 (2009), 439–454.

[23] Ganelson. 2025. Inform Website. https://ganelson.github.io/inform-website/. Accessed: 2025-01-30.

[24] David Gotz and Harry Stavropoulos. 2014. Decisionflow: Visual analytics for high-dimensional temporal event sequence data. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1783–1792.

[25] Craig Paul Green, Lars Erik Holmquist, and Steve Gibson. 2020. Towards the Emergent Theatre: A Novel Approach for Creating Live Emergent Narratives Using Finite State Machines. In *International Conference on Interactive Digital Storytelling*. Springer, 92–101.

[26] Melanie C Green and Keenan M Jenkins. 2014. Interactive narratives: Processes and outcomes in user-directed stories. *Journal of Communication* 64, 3 (2014), 479–500.

[27] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems* 36 (2023), 79081–79094.

[28] Shunan Guo, Ke Xu, Rongwen Zhao, David Gotz, Hongyuan Zha, and Nan Cao. 2017. Eventthread: Visual summarization and stage analysis of event sequence data. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 56–65.

[29] Inworld. 2025. Inworld: AI framework for building real-time agentic experiences. https://inworld.ai/. Accessed: 2025-01-30.

[30] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus Gross, and Hanspeter Pfister. 2017. Visualizing nonlinear narratives with story curves. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 595–604.

[31] Max Kreminski and John Joon Young Chung. 2024. Intent Elicitation in Mixed-Initiative Co-Creativity.. In *IUI Workshops*.

[32] Max Kreminski and Noah Wardrip-Fruin. 2018. Sketching a map of the storylets design space. In *Interactive Storytelling: 11th International Conference on Interactive Digital Storytelling, ICIDS 2018, Dublin, Ireland, December 5–8, 2018, Proceedings 11*. Springer, 160–164.

[33] Vikram Kumaran, Jonathan Rowe, and James Lester. 2024. NARRATIVEGENIE: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 20. 76–86.

[34] Ben Kybartas and Rafael Bidarra. 2016. A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games* 9, 3 (2016), 239–253.

[35] Michael Lebowitz. 1985. Story-telling as planning and learning. *Poetics* 14, 6 (1985), 483–502.

[36] Zhaohui Liang, Xiaoyu Zhang, Kevin Ma, Zhao Liu, Xipei Ren, Kosa Goucher-Lambert, and Can Liu. 2024. StoryDiffusion: How to Support UX Storyboarding with Generative-AI. *arXiv preprint arXiv:2407.07672* (2024).

[37] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[38] Shixia Liu, Yingcai Wu, Enxun Wei, Mengchen Liu, and Yang Liu. 2013. Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2436–2445.

[39] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv 2303.16634* (March 2023). https://www.microsoft.com/en-us/research/publication/gpteval-nlg-evaluation-using-gpt-4-with-better-human-alignment/

[40] Zhicheng Liu, Bernard Kerr, Mira Dontcheva, Justin Grover, Matthew Hoffman, and Alan Wilson. 2017. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 527–538.

[41] Qiang Lu, Xiang-yuan Zhu, Li Liu, and Shu-bo Cao. 2014. An effective demonstration for group collaboration based on storyline visualization technology. In *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 47–52.

[42] Robert McKee. 1997. Substance, structure, style, and the principles of screenwriting. *Alba Editorial* (1997).

[43] James R Meehan. 1977. TALE-SPIN, An Interactive Program that Writes Stories.. In *Ijcai*, Vol. 77. 91–98.

[44] Marco Mesiti, Mario Pennacchioni, and Paolo Perlasca. 2023. Indexing Structures for the Efficient Multi-Resolution Visualization of Big Graphs. *IEEE Access* (2023).

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

[46] Muhammad U Nasir, Steven James, and Julian Togelius. 2024. Word2World: Generating Stories and Worlds through Large Language Models. *arXiv preprint arXiv:2405.06686* (2024).

[47] Netflix. 2025. Black Mirror: Bandersnatch. https://en.wikipedia.org/wiki/Black_Mirror:_Bandersnatch. Accessed: 2025-02-06.

[48] Choice of Games. 2025. ChoiceScript Intro. https://www.choiceofgames.com/make-your-own-games/choicescript-intro/. Accessed: 2025-01-30.

[49] Michael Ogawa and Kwan-Liu Ma. 2010. Software evolution storylines. In *Proceedings of the 5th international symposium on Software visualization*. 35–42.

[50] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

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

[53] Xiangyu Peng, Jessica Quaye, Sudha Rao, Weijia Xu, Portia Botchway, Chris Brockett, Nebojsa Jojic, Gabriel DesGarennes, Ken Lobb, Michael Xu, et al. 2024. Player-driven emergence in llm-driven game narrative. In *2024 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[54] Steven Poulakos, Mubbasir Kapadia, Guido M Maiga, Fabio Zünd, Markus Gross, and Robert W Sumner. 2016. Evaluating accessible graphical interfaces for building story worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*. Springer, 184–196.

[55] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[56] Jay Ratican and James Hutson. 2024. Adaptive Worlds: Generative AI in Game Design and Future of Gaming, and Interactive Media. *ISRG Journal of Arts, Humanities and Social Sciences* 2, 5 (2024).

[57] Mark Owen Riedl and Vadim Bulitko. 2013. Interactive narrative: An intelligent systems approach. *Ai Magazine* 34, 1 (2013), 67–67.

[58] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.

[59] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[60] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[61] Emily Short. 2009. Moods in Conversation. https://emshort.blog/2009/12/10/moods-in-conversation/. Accessed: 2025-04-14.

[62] Ulrike Spierling and Nicolas Szilas. 2009. Authoring issues beyond tools. In *Interactive Storytelling: Second Joint International Conference on Interactive Digital Storytelling, ICIDS 2009, Guimarães, Portugal, December 9-11, 2009. Proceedings 2*. Springer, 50–61.

[63] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[64] Sangho Suh, Jian Zhao, and Edith Law. 2022. Codetoon: Story ideation, auto comic generation, and structure mapping for code-driven storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[65] Penny Sweetser. 2024. Large language models and video games: A preliminary scoping review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–8.

[66] Yuzuru Tanahashi, Chien-Hsin Hsueh, and Kwan-Liu Ma. 2015. An efficient framework for generating storyline visualizations from streaming data. *IEEE transactions on visualization and computer graphics* 21, 6 (2015), 730–742.

[67] Tan Tang, Renzhong Li, Xinke Wu, Shuhan Liu, Johannes Knittel, Steffen Koch, Thomas Ertl, Lingyun Yu, Peiran Ren, and Yingcai Wu. 2020. Plotthread: Creating expressive storyline visualizations using reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 294–303.

[68] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are Large Language Models Capable of Generating Human-Level Narratives? *arXiv preprint arXiv:2407.13248* (2024).

[69] Chao Tong, Richard C Roberts, Robert S Laramee, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, and Xiaojuan Ma. 2018. Storytelling and Visualization: A Survey.. In *VISIGRAPP (3: IVAPP)*. 212–224.

[70] Twine. 2025. Twine: An Open-Source Tool for Interactive Storytelling. https://twinery.org/. Accessed: 2025-01-30.

[71] Huafeng Wang, Xian Zhang, Yinxing Jin, and Xixin Ding. 2024. Examining the relationships between cognitive load, anxiety, and story continuation writing performance: a structural equation modeling approach. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–10.

[72] Yi Wang, Qian Zhou, and David Ledo. 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.

[73] Stephen Ware and R Young. 2011. Cpocl: A narrative planner supporting conflict. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 7. 97–102.

[74] Krist Wongsuphasawat and David Gotz. 2012. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2659–2668.

[75] XKCD. 2025. XKCD: A webcomic of romance, sarcasm, math, and language. https://xkcd.com/657/. Accessed: 2025-02-03.

[76] Qing Ru Yong and Alex Mitchell. 2023. From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *International Conference on Interactive Digital Storytelling*. Springer, 395–409.

[77] R Michael Young, Stephen G Ware, Brad A Cassell, and Justus Robertson. 2013. Plans and planning in narrative generation: a review of plan-based approaches

to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative* 37, 1-2 (2013), 41–64.

A Participant Information

Table 3: User study participant information including, enagagement type and years of experience

Participant ID	Engagament	Experience
P1	Hobby	1
P2	Hobby	15
P3	Professional	2
P4	Professional	7
P5	Hobby	8
P6	Hobby	10
P7	Hobby	8
P8	Hobby	1
P9	Hobby	10
P10	Hobby	did not answer
P11	did not answer	2