Toward Personalizable AI Node Graph Creative Writing Support: Insights on Preferences for Generative AI Features and Information Presentation Across Story Writing Processes

Hua Xuan Qin
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
hxqin682@connect.hkust-gz.edu.cn

Guangzhi Zhu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
gzhu305@connect.hkust-gz.edu.cn

Mingming Fan
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China
mingmingfan@ust.hk

Pan Hui
Computational Media and Arts
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
panhui@ust.hk

Abstract

As story writing requires diverse resources, a single system combining these resources could improve personalization. We leverage the broad capabilities of generative AI to support both more general story writing needs and an understudied but essential aspect: reflection on the moral (lesson) conveyed. Through a formative study (N=12), a user study (N=14), and external evaluation (N=19), we designed, implemented, then studied a prototype plugin for FigJam supporting visualization of the story structure through customizable node graph editing, LLM audience impersonation (chatbot and non-chatbot interfaces), and image and audio generative AI features. Our findings support writers' preference for leveraging unique interplays of our breadth of features to satisfy shifting needs across writing processes, from conveying a moral across audience groups to story writing in general. We discuss how our tool design and findings can inform model bias, personalized writing support, and visualization research.

CCS Concepts

• **Human-centered computing → Graphical user interfaces**; *Empirical studies in HCI*; • **Computing methodologies → Natural language processing**.

Keywords

Creativity Support, Writing Assistants, Visualization, Human-AI Collaboration

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713569

ACM Reference Format:
Hua Xuan Qin, Guangzhi Zhu, Mingming Fan, and Pan Hui. 2025. Toward Personalizable AI Node Graph Creative Writing Support: Insights on Preferences for Generative AI Features and Information Presentation Across Story Writing Processes. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3706598.3713569

1 Introduction

Recent works support the possibly superior capabilities of large language models (LLMs), such as Generative Pre-trained Transformer 4 (GPT-4) [96], in simultaneously satisfying diverse needs during story writing (creative writing works, such as novels): literary (e.g., character conversation), technical (e.g., grammar), and different levels of creative control (ownership) [36, 51, 108, 149, 158, 163]. Though none focuses on understanding how their tool could support reflection on the moral, the lesson (e.g., "be honest") explicitly stated [139] or implied [43], that a story might convey to its audience. Many believe that an essential purpose of stories is to convey morals that promote prosocial behavior [27, 59, 147]. Even when a story is not created with a moral in mind, the audience might extract an unintended one [53, 122]. Concerns for a misunderstood moral's impact on the audience and society [14, 19, 85, 93, 117, 118, 120, 152] warrant research on supporting writers' reflection on the morals potentially conveyed. As a single interface integrating different resources could improve the writing experience [70, 112], in line with shifting needs found across story writing processes [51, 108], a tool could support more diverse users by combining features supporting both reflection on a moral and other story writing needs, by supporting story writing around a moral, the writing of any story with consideration of the moral potentially conveyed.

Appreciation of the moral mainly depends on the understanding of written expression (e.g., vocabulary) and the logical relationships between story elements (e.g., events), which both can depend on the

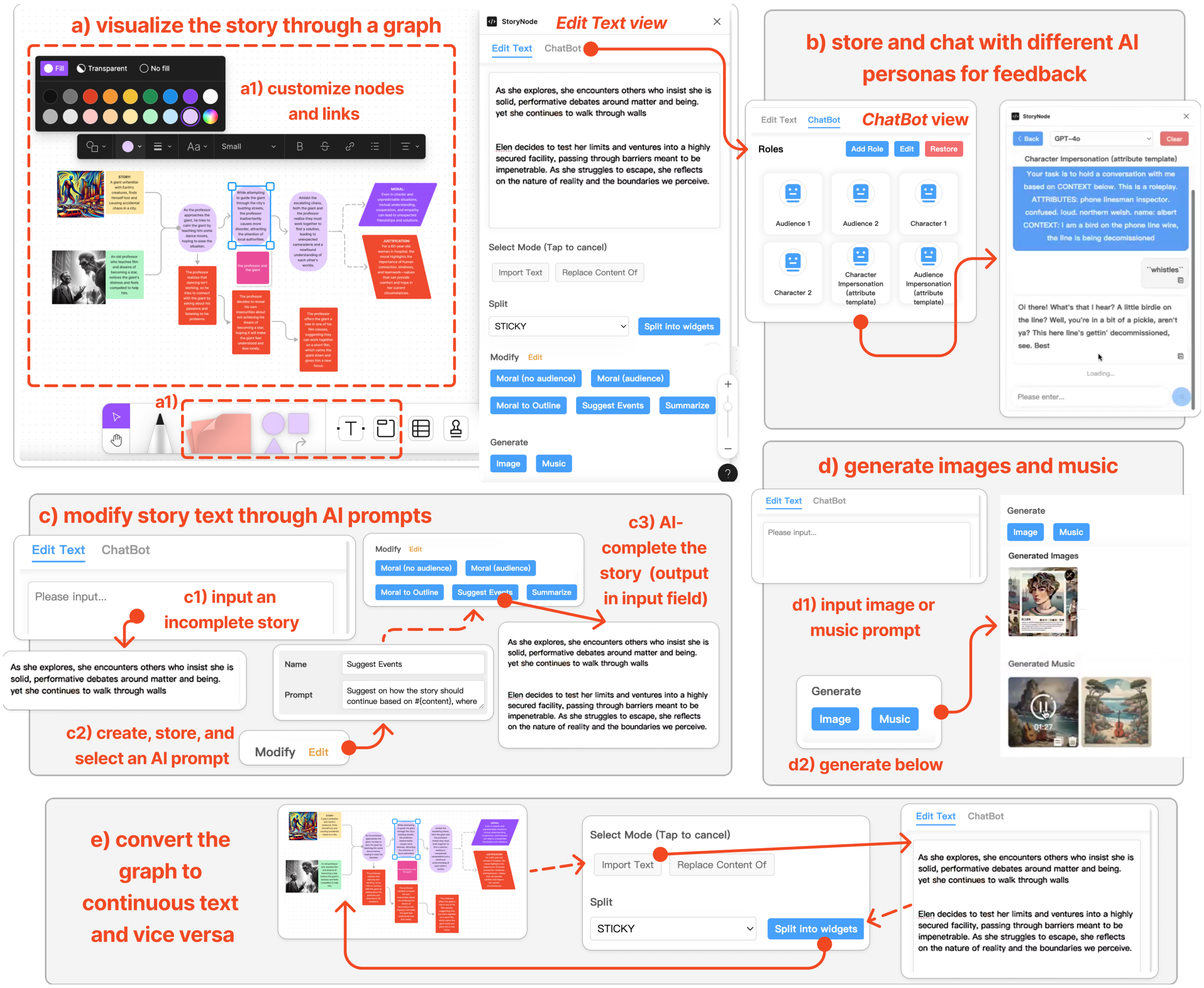

Figure 1: StoryNode, our plugin for FigJam [46] (right of a), combines different story writing resources into a single system to improve personalization for a broader user group. During a writer's block, the writer can obtain sources of inspiration through generative AI (GPT-4o, Dall-E 3, and Suno v3), such as character conversation (b), story completions (c), and images and music (d). When reviewing, they can obtain feedback from the audience's perspective (b and c). They can also visualize the plot structure by creating a node graph (a). To differentiate between types of information (e.g., story versions and character information), the writer can customize node and link appearance (a1) and insert generated text, image, or music. As a writer can shift between a continuous text format and a graph format, we also facilitate conversion (e).

audience's background [14, 94, 122, 144, 147, 157]. On top of providing the aforementioned story writing support, LLM can provide feedback reflective of the audience through impersonation [18]. To further complement writers' cognitive processes, other LLM storytelling support works have explored the addition of sources of inspiration beyond text, mainly image and audio, and interactive visualization of the plot (story event sequence) logic through a graph (e.g., [10, 36, 108, 110, 129, 155]). Both can support understanding of story relationships [110, 155]. Among graphs, node graphs can support more intuitive visualization of more diverse relationships through nodes representing different types of information and links, their relationships [155]. Though two improvements can be made to existing LLM-powered node graph storytelling support tools [110, 155]. Firstly, they require integrating nodes representing different sub-components of story events (e.g., character and action), which can affect clarity for more complex stories [110, 155]. Requiring nodes containing descriptions of events (event nodes) alone (i.e., an event node graph) can be enough to support the visualization of logic [74], facilitate exploration of story branches, and support conversion to the story text [31, 45]. Secondly, as data visualization research [16, 68, 79] suggests individualized preferences for colors and shapes, adding customization options for node/link colors and shapes (e.g., rectangular nodes and solid or dashed lines) could support visualization of more diverse relationships (e.g., event importance).

Experiment-wise, given the potential practical relevance of combining empirical evidence with theoretical knowledge expertise [102, 103], obtaining feedback from creative writers familiar with story writing could lead to design implications relevant to future AI technologies as well. Moreover, since interaction and story writing needs can be influenced by one's cultural background [51], a culturally diverse group could lead to more generalizable insights for more inclusive design [78]. Though existing LLM audience impersonation or graph editing works do not focus on story writing, creative writers, and/or their cultural diversity [18, 110, 155].

Our research question is thus: *how can visualization of the plot as a node graph augmented by node/link appearance customization, LLM impersonation, and image and audio generative AI features facilitate story writing around a moral?* Our approach is to design, implement, then evaluate a prototype system with creative writers of diverse cultural backgrounds. As illustrated in Figure 2, informed by a formative study and LLM model evaluation, we adapted FigJam [46], a popular online whiteboard platform, to generative AI-powered event node graph editing by developing StoryNode, a plugin with LLM-powered chatbot and non-chatbot interfaces and image and audio generation. Through a within-subject user study with 14 creative writers, we obtained insights on writers' thinking and usage patterns when using an LLM (FigJam/StoryNode) and a popular non-LLM tool. Through an external evaluation of task responses with 19 creative writers, we obtained insights from the perspective of the audience.

Key findings suggest that writers of diverse cultures can share the goal of conveying morals across cultures, leveraging interplays of our breath of features. They preferred a tool combining such features even for story writing in general, even more if the tool selectively shows features based on needs shifting across writing processes. While such series of feature needs can be unique, they can be grouped into higher-level factors reflecting writing and visualization theories.

Our contribution is thus threefold. First, we conducted a formative study with 12 creative writers, identifying design needs for leveraging graph editing and generative AI for story writing around a moral. Second, based on these needs, we designed StoryNode[1], a plugin integrating generative AI features with FigJam's default node/link customization. This design can readily be used for various other cases (e.g., academic writing or collaborative story writing) given FigJam's availability and support for collaboration. Third, through creative writer author (N=14) and evaluator (N=19) feedback and observations, we present the first findings on the interplay between factors that could influence usage patterns for a system combining customizable graph editing and chatbot and non-chatbot text, image, and audio generative AI features. These could inform writer profiles for the design of tools personalizable across AI technologies, culturally diverse writer or audience groups, and social dynamics (e.g., human-human collaboration) and cognitive process research on story visualization for story writing in general.

[1]Link to GitHub repository: https://github.com/ropenstick/StoryNodePlugin

2 Related Work

2.1 Definitions

We define stories as accounts of interconnected events (major changes) with real or imaginary actors, "characters" [7], and creative writing as the creation of original text-based works, such as novels, movie scripts, and interactive fiction [108]. As creative writing skills can be developed through various formal or informal means and there is no standard for assessing expertise, we follow prior user study works' example to recognize anyone who has authored creative writing works (e.g., stories) as a creative writer without categorizing their level of expertise [51, 108]. For reference, we report demographic information about creative writing experience.

2.2 Varied Needs for Story Writing Around a Moral

Prior work suggests that support needs found for story writing in general can fall across four main dimensions: 1) linearity of processes, 2) storytelling approaches, 3) sources of inspiration, and 4) audience. As stages within writing processes can build upon each other [47], the introduction of reflection on the moral can affect such needs to varying extents.

For 1), more general writing processes can be seen as possibly iterative series of planning (organizing ideas and goals), translating (turning ideas into writing), and reviewing (evaluating the work so far) in no particular order [47], with different types of AI support required for each (e.g., brainstorming and outlining for planning, vocabulary and story scene description for translating, and audience feedback for reviewing) [18, 51]. For conveying a moral, while some storytellers might prefer to start with planning, others prefer to start more "freely" [133], finding the message as they write [26]. Both system and study design should thus expect as much diversity in the linearity of story writing processes with reflection on a moral.

For 2), writers might adopt different storytelling approaches based on their focus on the key story elements, plot and character: on the causality between events for a plot-driven approach, the agency of characters for a character-driven one, or a balance [143, 146]. As a moral can be understood through the plot's progression or the evolution of the character [11, 114, 122, 146], support needs for either storytelling approach can be present.

For 3), sources beyond text, such as image and audio, can inspire written expression, by complementing writers' mental imagery [8, 40, 48, 80, 124, 146], mental representations (e.g., bits of story scene visuals [108]) not directly triggered by external stimulus [106]. Images can also serve as visual aids for understanding the plot progression, as seen with storyboards [30]. Thus, sources beyond text could be complementary to reflection on a moral.

For 4), compared to more personal forms of storytelling (e.g., personal journaling [69]), story writing around a moral places a greater focus on communicating values to an audience, whose understanding might differ from the writer's due to various, possibly

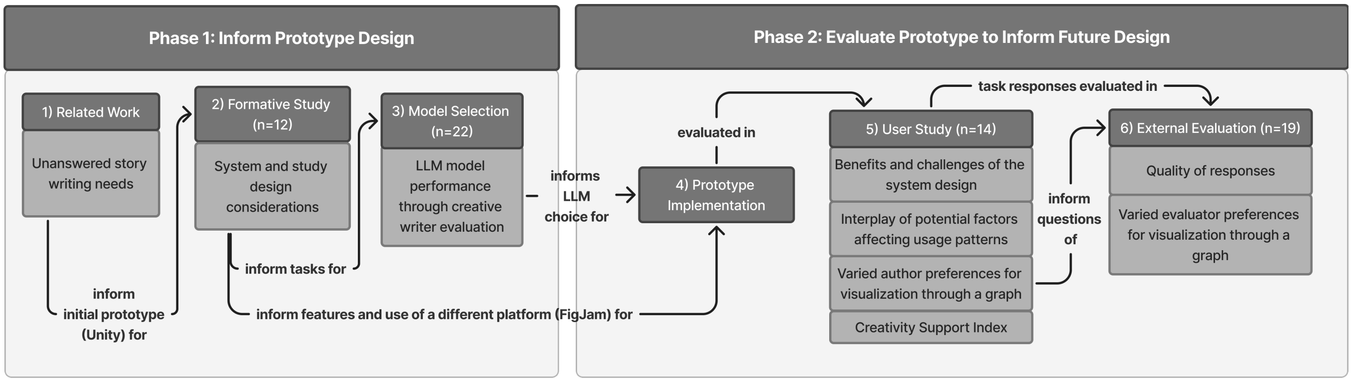

Figure 2: Overview of our work. We started by reviewing related work for story writing needs (1), which inspired the design of an initial prototype. This was evaluated through a formative study (2). Resulting design considerations informed our selection of a suitable LLM model (3). Both the design considerations and model selection evaluation informed the implementation of our final prototype (4). This was evaluated through a user study (5), where creative writers completed writing tasks with it. The task responses were evaluated through an external evaluation (6). The user study also led to findings on authors' varied visualization preferences (5). This inspired us to obtain preferences from a different perspective, that of evaluators (6).

overlapping factors: domain expertise, culture, and cognitive development, sometimes associated with age [17, 67, 92, 147]. In particular, moral judgment works have mainly focused on culture [54], finding that cultural differences, among geographic areas, religious beliefs, and income levels for instance [54], could influence prioritization of different values (e.g., individual rights versus societal obligations [25, 55, 56, 162]) and causal inferences (e.g., attributing one's behavior to personal traits versus contextual factors [141]). Works have also expressed concerns over (commonly used) LLMs' cultural biases in causing homogenization of the writing, the loss of the author's voice, perpetuation of stereotypes, and loss of cultural identity [3, 18, 141]. To understand potential resulting variations in needs, we attempt to obtain system design and evaluation feedback from a culturally diverse participant group.

2.3 LLM Capabilities for Story Writing Around a Moral

Many non-generative-AI works have focused on the generation of stories centered around specific ideas [7], including the moral of the story [1, 122]. LLM works have focused on moral reasoning capabilities across domains [15, 24, 44, 57, 66, 111, 123, 127], evaluation of underlying meanings [27, 35, 57, 75, 131, 132], generation of new story content [23, 27, 88, 101, 149], impersonation [18, 50, 82, 108], and summary generation [28, 29, 71, 107, 132], which depends on the ability to extract underlying meanings [145]. While some have focused on generating entire stories based on morals [84], with possibly superior performance in adapting to audience preferences [159] compared to non-LLMs [122], or providing feedback on morals [61, 145], none studies user interaction during story writing. Such findings with a tool supporting different levels of creative control (e.g., entire story generation to brainstorming) could inform the design of AI interfaces more reflective of differing views on AI contribution [21], potentially addressing recurrent concerns on homogenization [3, 51, 72].

2.4 AI and Graph Editing Creative Writing Support

AI creative writing support works have focused on visualization through node graphs [39, 110, 155], other visuals showing plot progression and/or character interaction over it [36, 62, 63, 86, 148], and visuals and/or audio relevant to specific events and characters [10, 34, 86, 108, 110, 129, 148, 155], with all generative AI works focused on LLMs. To observe processes of varying linearity, we are inspired by prior LLM creative writing support research [108, 110] to not limit our study to a specific stage (e.g., planning) nor the generative AI functionalities to specific prompts (i.e., by supporting freeform prompt input). To support different storytelling approaches, we complement plot visualization through an event node graph [4, 45, 58, 65, 73, 89, 90, 104, 113, 128, 130, 151] with character conversation through LLM impersonation, which has been shown to support both character and related plot construction [108, 124]. While node graph works have acknowledged shapes and colors' potential to augment story visualization by representing additional information [12, 31, 164] without overloading the viewer [2], they have rarely explored how and why preferences can vary among authors. While data visualization research not focused on 1) story writing or 2) event node graphs have suggested individualized preferences (e.g., [16, 22, 68, 79, 125, 140]), visualization needs can vary based on 1) contexts, such as school subjects [22] versus story element relationships, or 2) visual components available, such as the use of mainly line colors to represent story relationships in line graphs [140] instead of node and line colors. Studying potential factors influencing preferences for node/link appearance customization for event node graphs could further inform personalized story visualization. For sources of inspiration beyond text, we focus on a tool combining image and audio to accommodate users' technology availability [108]. We also leverage LLM impersonation to provide feedback from the audience's perspective, which has rarely been done even for writing in general [18].

3 Formative Study

We conducted a formative study to obtain additional empirical evidence 1) on how generative AI (e.g., character/audience impersonation, audio, and images) could make a creative writing support tool with event node graph editing features more personalizable (Section 3.4) and 2) on how writers would evaluate a story (Section 3.5), placing our focus on writing around a moral.

3.1 Participants and Procedure

Through word-of-mouth and social media, we recruited 12 creative writers (5 females and 7 males aged 19-33, average of 26.9; anonymized as F1, F2, ...), as described in Table 1. Through questions inspired by related work [108], we collected information on cultural, linguistic, creative writing work and education experience, and attitudes toward AI contribution, which can influence understanding and expression of a moral [14, 20] and story writing support needs [18, 51, 60]. While culture is made of various factors (Section 2.2), for comparability and less privacy concerns, we follow the example of existing AI writing support and LLM cultural bias works [3, 51, 108, 122, 141] to collect geographic locations (mainly countries) as proxies for cultures.

After being introduced to key concepts (i.e., moral of a story, event node graph, and generative AI), each participant completed a task through two different event node graph tools (2 tasks in total) for comparison: create the outline (list of plot event descriptions) for a story centered around a moral in an event node graph, where each event description is 1-2 sentences long. As there seems to be no close reference, we created our own task to balance participants' exploration of use cases for the tools' different features and their availability. Specifically, to accommodate potentially diverse writing processes (Section 2), we impose little restriction to the writing process (e.g., writing the story or the outline first), any other information included in the graph, time and word limits, and branching type (i.e., branching, where the story branches out to different versions, like in interactive fiction, or non-branching, where the story only has one final version, like in traditional novels). As plots can widely vary in lengths (e.g., novel versus short story), we do not enforce the creation of full stories. Instead, we leverage a combination of empirical evidence, theories, and speculation, which can also have practical relevance to design [102, 103]. Specifically, we ask participants to describe potential use cases for their entire processes.

Each participant was also shown prompt examples and tried different existing text, image, and audio generation platforms. Upon completion of the task, each participant shared their experience, online or in person, through semi-structured interview questions about creative writing ("According to you, what is a successful story with a moral? How would you measure that?") and the tool design ("What features of the tools did you find useful for your task?" and "How could generative AI features augment your use of the prototype?"). Our study design, including the task design, was first pilot-tested by 3 creative writers for the suitability of content and length. For the data analysis, interview sessions ranged from 20 minutes to 1 hour, with additional notes obtained afterward. Each participant was offered a compensation of about 4 USD.

3.2 Tools Studied

The event node graph tools studied are Twine [49] and a prototype created using resources for Unity [115]. Twine [49] is an open-source tool for creating hypertext fiction (branching narratives made through hyperlinks) through an event-node-graph-like editing interface. Each node opens to a text editor window where the author can mix natural language story text with code segments, to include hyperlinks to story branches for instance. This can support writers without much programming knowledge [45] but is the only custom way to link nodes. Given Twine's popularity in potential participants' communities and reported intuitiveness [38], we explored whether some of its features could be relevant to our design. To diversify findings, our prototype supports drag-and-drop interaction for linking nodes, common among diagram software (e.g., draw.io [81]), and leverages color and shape customization options (Figure 3). Originally, we intended to augment our prototype, but writers' feedback led to a plugin for an existing platform (Section 5).

3.3 Data Analysis

Interviews were voice-recorded, automatically transcribed, anonymized, then manually reviewed by the same researcher who conducted all sessions. Transcripts, written observations, and post-session notes were then analyzed using thematic analysis [37], an approach for finding patterns within qualitative data. We started with the themes of "system design needs" (Section 3.4) and "system evaluation needs" (Section 3.5), given the goals of this study (Section 3), but obtained sub-themes (e.g., "graph editing features") inductively. Specifically, two HCI researchers first reviewed all data then iteratively analyzed it independently and discussed to agree on codes and themes. For example, for "system design needs", the quote "[A] tool is more effective than another [if] it makes the writing process smoother graphically." was associated with the code "Need for intuitive interaction for a graph editing interface", under the sub-theme "graph editing features". For "system evaluation needs", the quote "Culture could be important. People from Eastern versus Western cultures could see things differently." was associated with the code "Audience's cultural experience as a factor affecting understanding of a story's moral".

As researchers' cultural and professional backgrounds could have influenced their data analysis, we disclose them for future reference. Both researchers hold computer science degrees, have leveraged generative AI for writing, and enjoy reading stories in different languages. One has grown up in a Western society, received story writing education (classes), had part-time story writing experience, and written stories in different languages. The other has grown up in an Eastern society and had entrepreneurial experience in generative AI interface development.

3.4 Findings on System Design

Participant feedback suggests varying needs for the following.

3.4.1 Graph Editing Features. Participants expressed individualized needs in customizing graph nodes and links, mainly shapes,

ID	Age	Gender	Countries	Languages	Professional Experience	Education	AI	At home?
F1	30	Female	CN	EN	FP	Formal (degree)	L	N
F2	33	Male	UK	EN	P	Formal (workshops)	S	Y
F3	31	Female	CN	CH	FP	Formal (degree)	S	Y
F4	28	Female	CN, UK	EN	P	Formal (classes)	S	Y
F5	24	Male	CN	CN, EN	P	Formal (degree)	L	Y
F6	28	Female	CN	CN	P	Informal	S	Y
F7	30	Male	CN	CH	FP	Formal (classes)	L	Y
F8	19	Male	India	A, EN	P	Informal	S	N
F9	25	Male	Canada, CN	EN, CH	P	Informal	S	Y
F10	23	Male	CN	CN	FP	Formal (classes)	E	Y
F11	26	Female	Canada	EN, FR	None	Informal	E	Y
F12	26	Male	Philippines	EN	P	Formal (classes)	S	Y

Table 1: Demographic information of participants from the formative study. For "Countries", geographical areas whose cultures the participants "spent more time living, working, and/or studying with", "CN" means *China* and "UK" *United Kingdom of Great Britain and Northern Ireland*. Given possible cultural and thus LLM bias difference among the different parts of China [141] (e.g., Mainland and Hong Kong), we record participants' specifications when available. For "Languages" the participants "usually write creative writing works" in, "EN" means *English*, "CH" *Chinese*, "A" *Assamese*, and "FR" *French*. For "Professional Experience", "F" means that the participant has had *"full-time professional experience in story writing"*, "P" *part-time*, and "FP" *both*. "Education" refers to education, formal or informal (e.g., self-learning), "related to story writing". For "AI", "S" means that, plagiarism concerns aside, the participant would "still feel like [they] are the author" with AI generating *corrections, improvements, or small parts they have difficulty with only*, "L" *large parts*, and "E" *the entire work based on a detailed outline*. "Y" in "At home?" means that the participant completed the tasks 'at home' due to availability, sharing screenshots of their works instead of having a live session with a moderator.

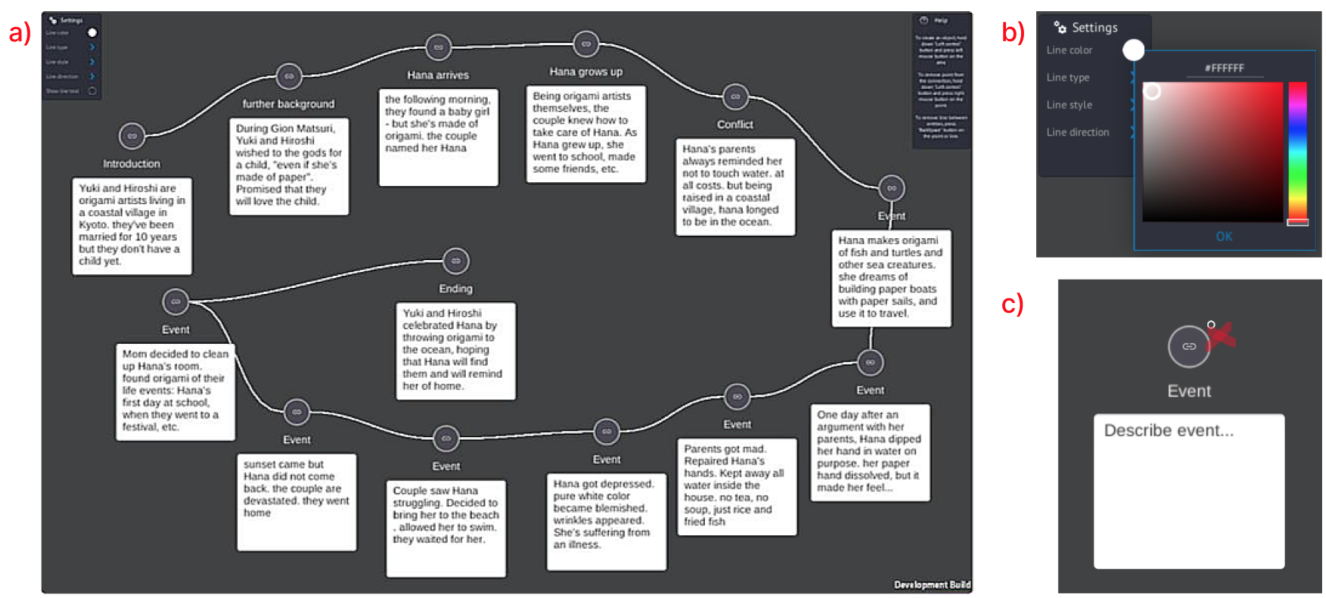

Figure 3: Examples showing our formative study prototype's GUI: a) the only view showing a formative study participant's event node graph, b) a zoomed-in shot of the link color and shape (e.g., solid or dashed) customization menu at the top left corner of a), and c) a larger image of a node, whose shape and input fields are inspired by those of Twine (i.e., title at "Event" and description). One can click anywhere to create a node then link it to another by hovering the mouse at the red cross in c) then dragging out a link.

colors, and compositions (graph components within nodes), to distinguish different types of information (e.g., excitement level of the event and character information). They also emphasized intuitive interaction. While most preferred the prototype (Figure 4), they

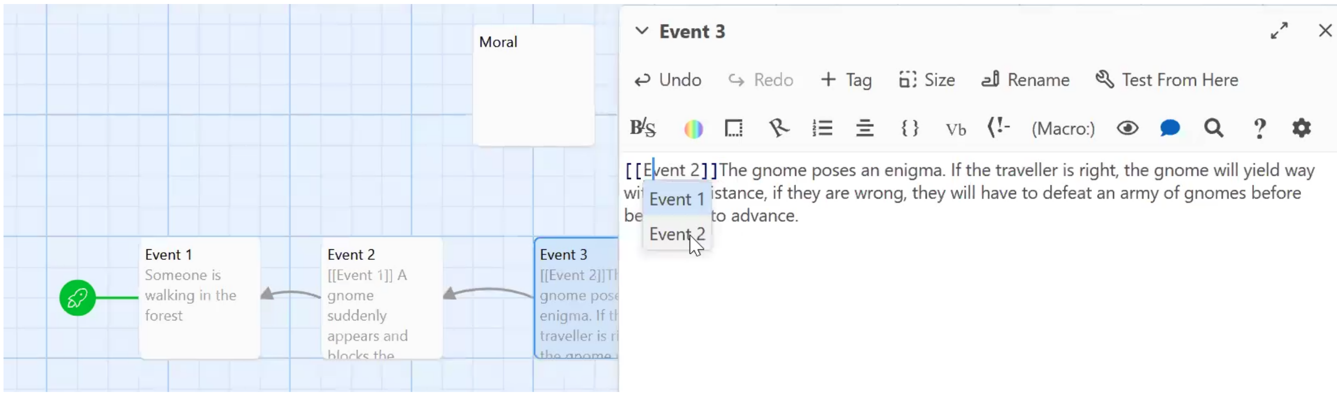

Figure 4: Screenshot of a participant recording partially showing Twine's interface. As seen on the right, to link two nodes, the user needs to type a code similar to "[[Event 2]]", where the name of the node needs to be inside the double brackets. Participants from our formative study preferred our prototype's "drag-and-drop" way for linking nodes over that of Twine as they found it more "intuitive", less "complicated", and "easier to get".

mentioned the "small hotspot area" (F1) for linking the nodes (Figure 3c) and missing zoom-in/out functionalities. They preferred a more "polished" platform, which can affect their "motivation" (F8). F4 suggested FigJam, which received unanimous preference from the nine who replied.

3.4.2 Generative AI Features. Participants agreed that AI generation features should produce output directly insertable into the graph. Though they described diverse use cases. LLM use cases mainly include feedback related to the moral based on the plot so far through audience or character impersonation, brainstorming, and technical writing support (e.g., grammar). For impersonation, participants are divided between wanting a chatbot interface, as it is more "immersive" for character conversation (F5), and an interface showing only suggestions at specific points of a story. While some participants believed that generated images and audio could augment their visualization of a graph, others agreed that they could be distracting. Though even these participants would like images and audio generation to augment visualization of an event when it is in "full screen", when the graph cannot be seen. Participants agreed that examples, templates, and prompts (e.g., character personas) would be useful.

3.4.3 Graph and Continuous Text Formats. While F5 preferred continuous text format only, others described alternating between continuous text and graph (e.g., for reviewing versus visualizing story relationships), envisioning features supporting such conversion.

3.5 Findings on System Evaluation

Most participants assessed how well the moral is conveyed and how well their story is written in general based on their audience's preferences. Commonly mentioned factors affecting audience's preferences include general life experience, cultural experience, and usual story consumption preferences. Others include age, education, and domain expertise.

3.6 Design Considerations

Our findings lead to 3 design considerations, all relevant to prior findings related to personalization. First, in line with prior findings on individualized preferences for data visualization [16, 22, 68, 79, 125, 140], **D1)** such a system should support diverse customization options for graph nodes and links through intuitive interaction in a "polished" GUI (Section 3.4.1). Second, in line with findings on sources of inspiration beyond text, on different use cases between a chatbot and a non-chatbot word-processor-like LLM interfaces, and on improved writing experience for an interface supporting the integration of generative AI output into the content being worked on [70, 108, 110, 112, 129, 155], **D2)** a system should integrate personalizable writing and visualization support through text (chatbot and non-chatbot interfaces), images, and audio (Section 3.4.2). Third, in line with needs to iterate between a graph and the story in continuous text [110, 155], **D3)** a system should support both graph and continuous text formats (Section 3.4.3).

For study design, based on formative study participant feedback, observations of diverse usage patterns, and potential participant availability, we also decided to focus on non-branching narratives and set the length of the event node graph to 5-10 events for the main story version that would be used for evaluation (Section 6.1.2). To include views on writing and reading preferences affecting story appreciation and broader definitions of cultural experience (Section 3.5), we started asking participants to report countries whose cultures they "like reading and/or writing stories about" in addition to ones whose culture(s) they "spent more time living, working, and/or studying with" (Section 3.1).

4 LLM Selection

To inspire participants with the latest advances, we aim to choose an LLM performing at least similarly to others in support for writing, impersonation, and outline creation around a moral needs (Section 3). As other functionalities have been explored (Section 2), we evaluated LLMs for the last by recruiting 22 creative writer evaluators of diverse cultural and creative writing experience (Table 6) to each answer a multiple choice questionnaire (Qualtrics [109]) comparing outputs of potentially top-performing LLMs, GPT-4o [98] and Claude 3 Opus [9], and human creative writers (as reference). Referring to prior work, formative study participants, and 3 generative AI creative support researchers, we asked evaluators to compare outputs between condition pairs (i.e., GPT/Claude, GPT/human, and Claude/human) for two tasks: 1) moral extraction (i.e., choosing the moral "that better corresponds to [a given] story outline") and 2) story outline generation based on a moral (i.e., choosing the "better" outline). To diminish biases caused by LLM training data, for the "[given] story outline[s]" and morals, we used 9 pairs of "yet-to-be-published" story outlines (5-10 events; average of 133 words) and morals covering diverse genres (e.g., fantasy, science fiction, realistic fiction/coming-of-age, horror, and mystery). For 2), instead of asking LLMs to generate entirely new stories, we asked them to modify the outlines as writers might prefer AI generation that follow their own stories' initial settings [21, 51] (Section A.2). Given similarity (Section A.3), **we chose GPT-4o**. Compensation was about 7 USD per evaluator and "yet-to-be-published" work authors.

5 System Design

Given unanimous preferences (Section 3.4.1), we leverage FigJam's interface and graph editing features (D1; Section 5.1) and extend it through a plugin (StoryNode) with generative AI (D2; Sections 5.2 and 5.3) and format conversion (D3; Section 5.2) features (Figure 1). We designed StoryNode's interface (e.g., input fields, prompt storage, and templates) based on prior generative AI writing support tool design works suggesting needs for greater freedom in designing AI prompts and for tracking information [18, 108, 110, 112] and formative study participants' needs for templates (Section 3.4.2; Figure 5). From a technical perspective, as parallel processing could better facilitate collaboration between human and generative AI for possibly long generation times [10], we ensured that text (GPT-4o [98]), image (Dall-E 3 [97]), and audio (Suno v3 [135]) generation can be requested in parallel by using different API keys. StoryNode was developed in two weeks in summer 2024, mainly with TypeScript and resources provided by the developers of FigJam [42].

5.1 FigJam *Whiteboard* View

The user opens to FigJam's whiteboard with default features for navigation (e.g., zoom in or out), drag-and-drop interaction for adding or resizing nodes/links, and node/link color and shape customization through a toolbar and/or by selecting a node/link (Figure 1a; **D1**).

5.2 StoryNode's *Edit Text* View

The user can then open StoryNode (a draggable window), to its *Edit Text* view. The input field (Figure 1c) supports a continuous text format, reminiscent of common word processor interfaces. The input field can be used for text, image, and audio generation (Figure 1c and d; **D2**). Text generation is triggered through a default or user-created button (Figure 1c) that will send a prompt, possibly containing a pointer to the input field content (Figure 5). Generated output will then replace the input field content. The user could start by writing some story content (e.g., events) in the input field, create and store a button through "Modify - Edit" (Figure 5) then click it to modify the input field content during a writer's block (e.g., story completion appealing to an audience group through impersonation) or review (e.g., correct grammar or obtain audience feedback on the moral conveyed). By selecting "Replace Content Of" then a node (Figure 1e), the user could 'store' input or output in it, which could also seem like the continuous text interface of a word processor when zoomed in (Figure 6a). The user could also convert graph to continuous text story content and vice versa (Figure 1e; **D3**) respectively by pressing "Import Text" then selecting nodes whose content will appear in ordered paragraphs in the input field (e.g., event node graph to outline) and by selecting a node shape under "Split", pressing "Split into widgets", then obtaining an ordered row of nodes containing the input field content split based on paragraphs (Figure 6c). For continuous text stored in a single node, they could first import it into the input field. Image and music generation takes in only the input field content (e.g., the story so far or a specialized prompt) then generate image and music files below (Figure 1d), which can all be inserted into the *Whiteboard* view's graph. For audio, we leverage an existing widget, a plugin for which many instances can be inserted into the whiteboard like a node, that takes in the URL to the generated audio file and creates a button for playing it [91] (Figure 6b). To mediate participants' varying needs for image and audio, we show user study participants (Section 6) keyboard shortcuts to zoom in or out unto a specific group of node(s), image(s), and audio file(s) to mimic visualization in "full screen" (Section 3.4.2).

5.3 StoryNode's *ChatBot* View

By selecting "ChatBot" in the navigation menu, the user can switch to its view (Figure 1b), where they can create different chatbot (**D2**) personas (e.g., characters, audience groups, and standard chat with an empty "Role Prompt"; Figure 5) and talk to them individually, with the conversation history saved until manually cleared.

6 User Study

To study usage and thinking patterns, we conducted a within-subject user study (pilot-tested by 4 creative writers of varying AI familiarity) in less than two weeks (Section 6.1) with 14 creative writers of diverse experience (Section 6.2) and had task responses evaluated by 19 creative writers in about two weeks (Section 7.5) in summer 2024.

6.1 Procedure

We planned a single user study session for each participant as follows and as shown in Figure 7.

6.1.1 Introduction. The moderator introduces the research goals and definitions with diverse examples (e.g., classic fables and original samples from Section 4 for the moral of the story and blogs and research papers for AI prompts).

6.1.2 Writing Tasks. To cover a broader range of potential usage patterns, we study Twine's event node creation and linking features (no AI) alongside the FigJam/StoryNode features mentioned in Section 5. For each condition, the participant needs to create an

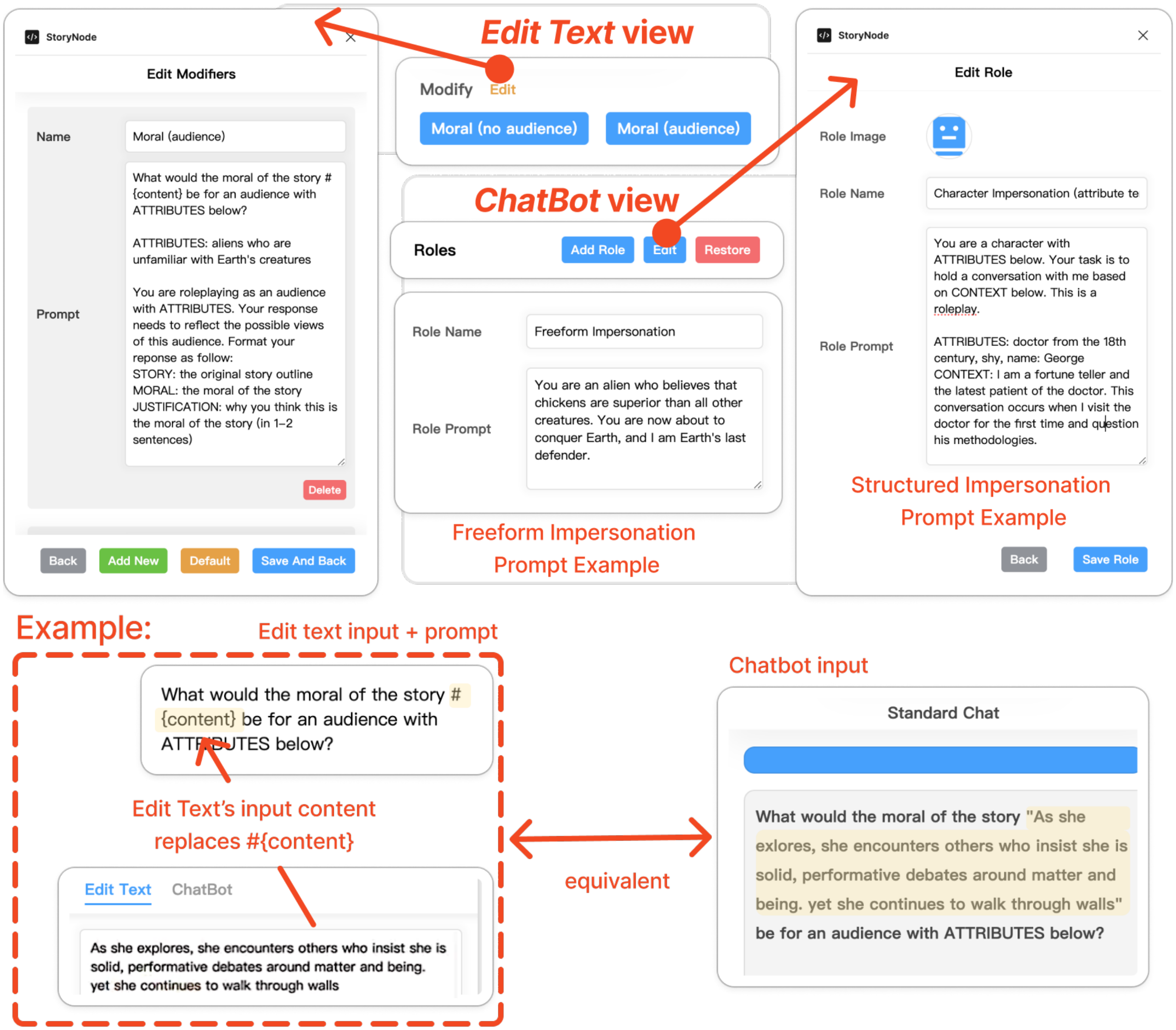

Figure 5: Screenshots of StoryNode showing input fields for storing a prompt for a non-chatbot interface (top left; accessed by pressing "Edit" in the *Edit Text* view; Figure 1c) and a chatbot persona (top right; accessed through "Edit" in the *ChatBot* view; Figure 1b). As illustrated (bottom), pressing a prompt button will replace any "#{content}" in its "Prompt" ("What would the moral of the story #{content} be...") by *Edit Text*'s input field content ("As she explores, she encounters..."). This is equivalent to entering a concatenation ("What would the moral of the story "As she explores, she encounters..." be...") into a chatbot. With only "#{content}" in "Prompt", the input field content can serve as the sole input for prompts used less often. Inspired by prior work and formative study participants, StoryNode comes with example prompts (e.g., audience feedback and event suggestion), and personas (middle top for freeform description and top right for structured based on attributes and context [18, 108]).

event node graph with 5-10 events for the main story version of a non-branching narrative with no other restriction, including on the creation of story versions other than the main. To diminish impact on writing processes, we let each participant complete two pairs of Twine-FigJam/StoryNode conditions with the order counterbalanced (Figure 7). To diminish biases related to personal preferences or familiarity for the external evaluation of responses (Section 7.5), we ask each participant to use the same audience group and moral of the story for each pair of Twine-FigJam/StoryNode conditions. Before the first task, the moderator introduces studied system features with examples then lets the participant explore. During the tasks, the participant is encouraged to think aloud.

Figure 6: Screenshots showing different system features: a) a zoomed-in node with FigJam's default features for changing font type, size, and style (e.g., bold) among others (dashed line in the figure), which are reminiscent of a common word processor interface, b) the widget that can be used like a play button for generated audio with a user-defined name (e.g., "Existential crisis") and connected to other nodes through links, and c) the input/output for "Split into widgets", which can facilitate conversion of a continuous text outline to an event node graph by taking in content in *Edit Text*'s input field then creating a row of ordered nodes, each containing a paragraph, in the *Whiteboard* view. a) and b) are from participant responses.

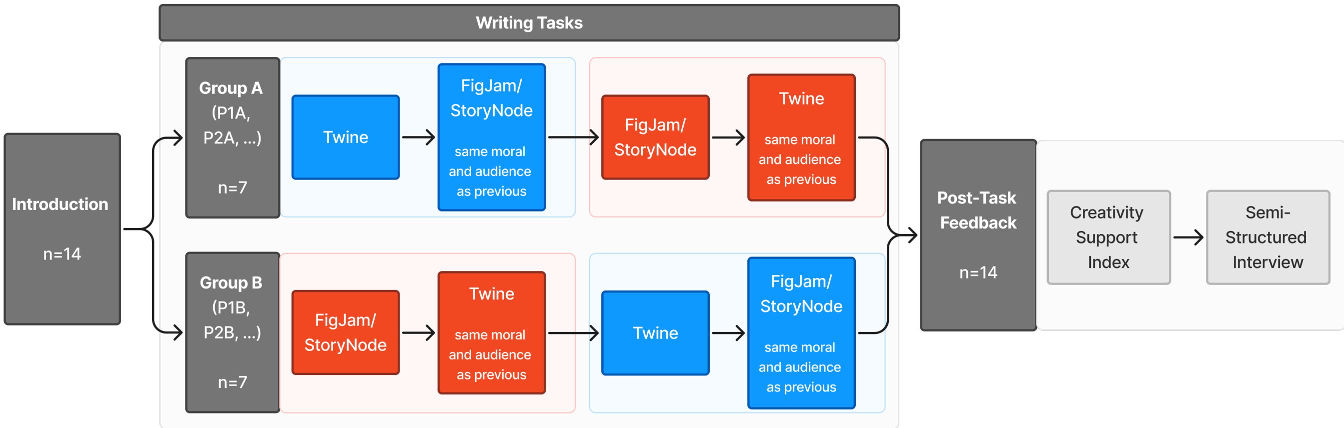

Figure 7: Flow of the user study. All participants begin with the introduction. For the writing tasks, half (P1A, P2A, ...) start with a Twine condition where they come up with a moral and an audience then complete a FigJam/StoryNode condition with the same moral and audience. Similarly, they then complete a FigJam/StoryNode condition and an ensuing Twine condition with the same moral and audience. The other half of the participants (P1B, P2B, ...) complete the same condition pairs in reverse order: FigJam/StoryNode-Twine then Twine-FigJam/StoryNode. After the tasks, all participants complete CSI questionnaires then semi-structured interviews.

6.1.3 Post-Task Feedback. After all writing tasks, the participant first completes a pair of Creativity Support Index (CSI) questionnaires, one for Twine and one for FigJam/StoryNode (Figure 10), then a semi-structured interview (Table 2). A CSI questionnaire is a survey that yields a CSI score (out of 100), which reflects a creative support tool's capabilities in supporting creativity for a specific creative task the respondent participated in. For each of the factors Expressiveness, Exploration, Enjoyment, Results Worth Effort, Immersion, and Collaboration, the user is presented two agreement statements (e.g., "I was able to be very creative while doing the activity inside this system or tool." and " The system or tool allowed me to be very expressive." for "Expressiveness") on a 10-point Likert scale, from "Highly Disagree" (1) to "Highly Agree" (10). The sum of each rating pair is then multiplied with the user's ranking of the corresponding factor's importance for the task being completed. This importance score is the number of times the user chose the factor in a series of pairwise comparisons between all factors for the statement "When doing this task, it's most important that I'm able to..." (higher score for more importance). The sum of all products divided by 3 is the CSI score [32]. Each participant thus completed the factor ranking only once in total. We use CSI scores to complement qualitative feedback and serve as future references.

6.2 Participants

Through word-of-mouth and social media, we recruited 14 creative writers (6 females and 8 males aged 19-33, average of 27.6), as described in Table 3. Each participant was compensated about 56 USD. Excluding breaks, 9 participants completed the experiment in one session (day), in about 3-4 hours. The rest split it over several days: about 1 (before third task) and 4 hours for P2A, 1 (before second task) and 5 hours for P3A and P4A, 2.5 (interview partially done) and 0.5 hour for P5A, and 3 (before interview) and 1 hour for P7B. Participants took about 20 minutes to over 1 hour per task and 0.5-1.5 hour for the interview. One researcher conducted all sessions remotely (through video call with StoryNode sent to the participant) or in person.

6.3 Data Analysis

Two researchers (Section 3.3) analyzed qualitative data, including interview transcripts, written observations, and post-session notes. Inspired by frequent suggestions of a system personalizing to user preferences based on profiles (Section 7.1.4), they associated participants' use cases and justifications to factors grounded in theory for practical relevance [102, 103], adopting both inductive and deductive thematic analysis strategies (e.g., [6, 112, 126, 142]).

Specifically, the researchers first agreed on the themes "potential factors influencing usage patterns" and "potential factors influencing preferences for information presentation in an event node graph". After reviewing writing process and data visualization works (Sections 2.2 and 2.4), they iterated between individually coding, inductively to find sub-themes (factors) within the two themes and possible new theme(s) (i.e., Section 7.1), and discussing to reach consensus. For "potential factors influencing usage patterns", as shown in Table 4, they ultimately grouped codes based on both factors and feature-specific sub-themes for more comprehensive insights on personalization to shifting needs across the writing process [47, 51, 60], obtaining the final theme of "potential factors influencing usage patterns across writing processes". The two themes on potential factors are thus made of both inductive and deductive insights, with "story length" (Section 7.3.3) derived inductively for information presentation.

7 Findings

We identified three themes from participant feedback: 1) benefits and challenges of similar systems (Section 7.1), 2) potential factors influencing usage patterns (Section 7.2), and 3) potential factors

Category	Questions
experience during the tasks	"How did you use different features during your writing tasks?"
potential use cases	"What features do you think would be helpful for your entire story writing process? Why?"
information presentation preferences	"During the writing tasks, what node/link color(s)/shape(s) have you used? Why? How about for a longer story?"
perceptions on impact	"(How) do you think the moral(s) conveyed through a story can affect the audience and society as a whole?" and "(How) do you think using generative AI to support the creation of stories centered around a moral can affect society?"

Table 2: Sample questions for the semi-structured interview of the user study (Section 6.1.3).

ID	Age	Gender	Locations	Professional Experience	Education
P1A	29	Male	CN*, Japan, UK, US	FP	Formal (degree)
P1B	27	Male	CN*, Germany, Japan, Singapore, UK, US	P	Formal (classes)
P2A	26	Male	Japan, Philippines, US	P	Formal (classes)
P2B	32	Female	CN, CN (HK), Japan, US	FP	Formal (degree)
P3A	28	Female	CN, Japan, South Korea	P	Informal
P3B	30	Female	CN, CN (HK)	FP	Formal (workshops)
P4A	19	Male	CN, France, US	P	Informal
P4B	30	Female	CN, CN (HK), Spain, Thailand	FP	Formal (degree)
P5A	26	Female	Canada	None	Informal
P5B	33	Male	UK	P	Formal (classes)
P6A	29	Male	CN, CN (HK), Iceland, New Zealand, UK	FP	Formal (workshops)
P6B	25	Male	CN, Russian Federation, US	P	Formal (degree)
P7A	25	Male	Canada*, CN	P	Informal
P7B	28	Female	CN	FP	Informal

Table 3: Demographic information of participants from the user study. For "Locations", geographical areas whose cultures the participants "like reading and/or writing stories about" or "spent more time living, working, and/or studying with", "CN" means *China*, "CN (HK)" *Hong Kong (S.A.R. China)*, "UK" *United Kingdom of Great Britain and Northern Ireland*, and "US" *United States of America*. An asterisk ("*") indicates the location whose culture the participant specified they have been most influenced by for the responses. Similarly to Table 1, we record the different parts of China. We use the same abbreviations for "Professional Experience" and "Education".

Quote	Code	Factor Sub-Theme	Feature-Specific Sub-Theme
"[M]aybe if you're writing character conversations, you'd use the chatbot. [...] I'm plot-driven. I prefer Edit Text."	Preference for a non-chatbot interface due to a plot-driven approach	storytelling approach	Varied use cases for a chatbot versus a non-chatbot interfaces across the writing process
"I don't really use conversation style for writing the outline. I guess the chat one would be more helpful if you're writing the actual story. [...] You can use the conversation with a chatbot to write the dialogues for your story."	Use of a chatbot interface for a lower level of abstraction	level of abstraction	

Table 4: A table showing examples for the thematic analysis of the user study. In the table, the two codes corresponding to different factor sub-themes can be grouped under a single feature-specific sub-theme. The researchers realized that, when viewed through both factors and feature-specific sub-themes, quotes from the same participant could provide more comprehensive insights on the interplay between factors or needs across the same writer's process [47, 51, 60].

influencing preferences for information presentation (Section 7.3). In line with prior work (Section 2), while participants' usual and experiment writing processes are generally described as iterative, stages within such iterations can roughly be categorized as planning, translating, and reviewing - which we refer to.

7.1 Benefits and Challenges

We identified three types of benefits: for conveying a moral (Section 7.1.1), for productivity (Section 7.1.2), and for various work types (Section 7.1.3). We also identified challenges (Section 7.1.4).

7.1.1 Support in Conveying a Moral. All participants mentioned the potential of a successfully conveyed moral in improving empathy, promoting prosocial behavior, and/or shaping society. Referring to their audience choices, not narrowed down based on culture (Table 8), 10 participants mentioned the goal of reaching an audience across cultures (e.g., "An effective story with a moral is universal." P2A and "I want it to be for a broader audience." P7B). All mentioned the potential of LLM impersonation in bridging the gap between the author and the audience when conveying a moral, with 13 focused on cultural differences. For instance, P5B explained, "It's hard to step outside my own cultural upbringing [...] GPT could be a good collaborator in this way to tell me 'insider' knowledge about a topic from another cultural perspective like you'd get working with another person, but it could do it for lots of perspectives simultaneously." For use cases, 3 participants specified creating personalized story versions to convey the same moral to different audience groups. All, including those who reported not usually using graphs (5 participants; e.g., Figure 8) or obtaining feedback from others (4 participants), observed or described iterating between graph editing and obtaining LLM feedback, mainly through impersonation. They found the former complementary to their exploration and review of plot logic, which they considered essential for reflection on a moral. While 13 mentioned cultural nuances in written expression (e.g., symbolism and vocabulary), participants were more divided for logic, with 3 believing in cultural differences and 3, not.

7.1.2 Support for Productivity. All participants agreed that a single system integrating customizable graph and AI (i.e., text, image, and audio) features could improve productivity for story writing in general, by saving time and costs spent trying to access them individually, especially given that different stories' writing might require different features (as elaborated by P1A and P1B). All participants preferred, in order, event node graphs, continuous text outlines, and full stories for ease of communicating story relationships (e.g., during review) and/or improving motivation by being less "tiring" to look at (quoting P1B and P7B).

7.1.3 Support for Various Work Types. 11 participants agreed that an event node graph would allow one to more clearly visualize different branching, which could support the creation of branching narratives (e.g., "RPG games" quoting P1B) and even non-branching ones, by supporting exploration of alternative story versions and morals (e.g., Figure 8). 10 participants mentioned that a graph editing system with image and audio generation features could facilitate storyboard creation for diverse works, such as comics, movies, visual arts, academic works, and application interface.

7.1.4 Challenges and Opportunities. We identified three main challenges. Firstly, while all participants preferred FigJam/StoryNode's availability of AI and graph editing features and interaction (e.g., linking nodes), they preferred Twine's "simpler" interface for specific stages. 11 participants found that seeing all features at once like in FigJam/StoryNode could be "distracting" (e.g., "I was not sure which feature I should start with." P7B). For this, 8 (of the 11) participants recommended various system features that could automatically recommend features and prompts, possibly by taking in a user profile as input to plan what the user needs at a specific event. Secondly, for graph visualization, 3 participants expressed nuances depending on the goal of the story writing and the type of stories. P6B maintained that, if one is "more focused on literary expression", giving the full story for review is always "better". P6A believed that an outline or graph is more suitable for stories dependent on "logic", such as those around a moral and detective stories. Similarly, P3A believed that an outline or graph might be less suitable for stories more focused on "emotions", like "high school romance". Thirdly, 6 participants expressed concerns over over-reliance on AI (not fine-tuned) for feedback on the moral of a story, 8 on cultural biases and 3 on less individualized works, which could better reach the general public but could harm its individuality. Though no participant mentioned any specific instance of cultural bias for the tasks (e.g., "I haven't encountered it yet." P5B). Despite 13 participants admitting the relevance to their writing goals, no participant requested LLM feedback through prompts describing the cultural backgrounds of audience groups (e.g., "[This] would require too much effort." P7B).

7.2 Potential Factors Influencing Usage Patterns Across Writing Processes

Participants described or were observed engaging in different use cases for different features and formats (i.e., graph versus continuous text) not only across their own writing processes but also between each other for similar stages of writing processes. Despite the diversity in usage patterns across these stages, participants' explanations mainly fall under 4 potential factors: storytelling approach (plot-driven, character-driven, or a balance), level of abstraction (focus on a lower or a higher level, respectively from details, e.g., event or character background information or passages in the story text, to relationships, e.g., between events and characters, or the overall impression of the story), motivation (practical reasons, such as diminished efforts and perceived practical constraints of specific circumstances), and mental imagery types (thoughts reported to be text and/or image for all participants) and clarity (clear or vague thoughts). To illustrate the interplay between these factors for a single writer's process, we focus on a specific participant (P2A)'s story writing process description, which more comprehensively represents other use cases (Section 7.2.1). To illustrate how writing processes can differ, we then compare P2A's process to others (Section 7.2.2).

7.2.1 Example of Interplay. **1) Chatbot versus non-chatbot (Edit Text) interfaces:** for LLM use cases, P2A justified their preference for a non-chatbot interface because they usually adopt a plot-driven approach. They saw a chatbot interface as more suitable for "writing character conversations" (storytelling approach). Though they mentioned including this use case for when they write specific

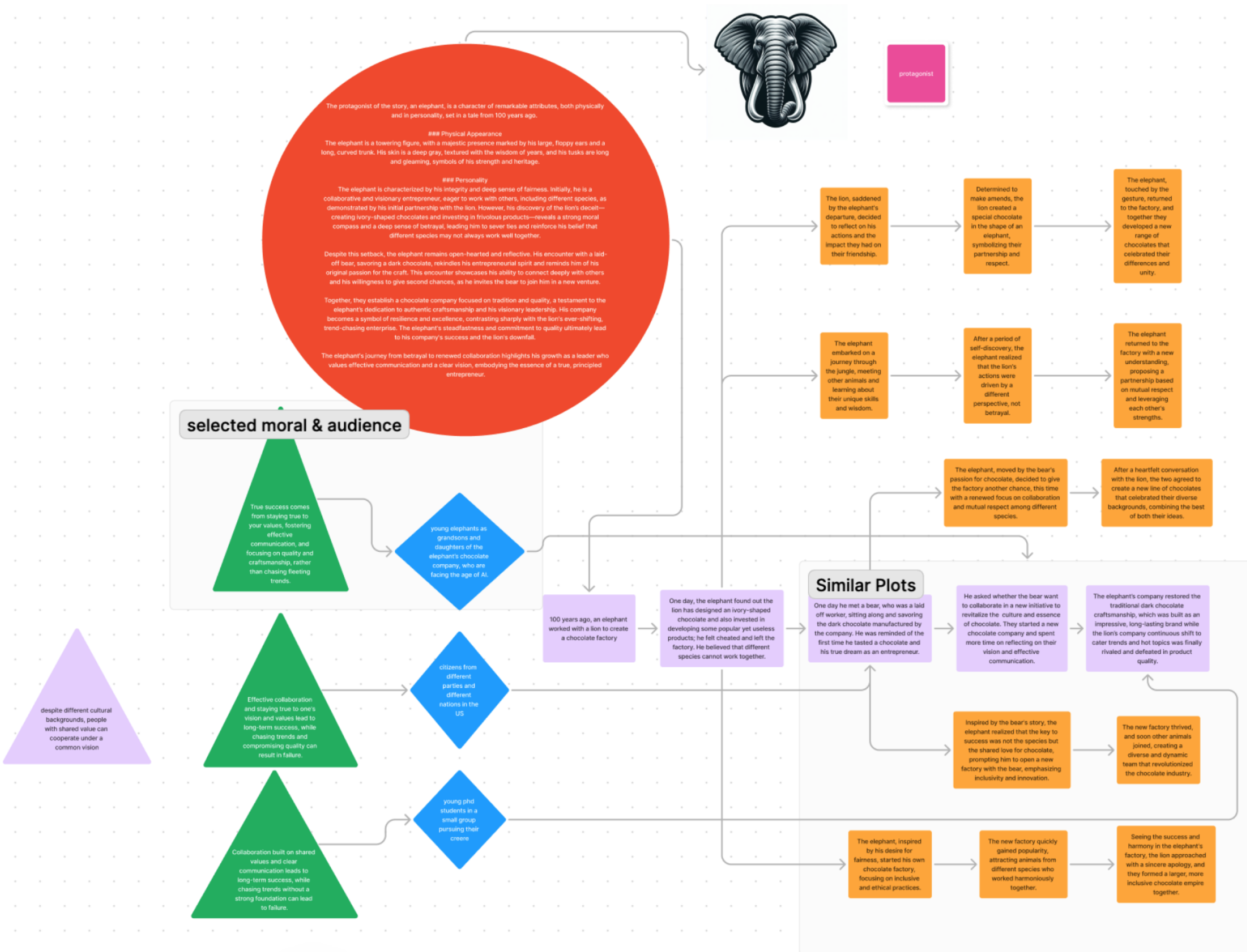

Figure 8: Screenshot of one of P6B's task responses. While they reported usually only using continuous text for story writing, they described preferring a "node graph" for visualization during review of "alternative storylines", "to clearly see the events, and whether it's clear to the audience". Similarly, their graph shows various plot versions and relationships between events (rectangular nodes), morals (triangles), and audience (diamonds, circle, elephant image, and music button at its right).

scenes in their story (lower level of abstraction). **2) AI contribution:** P2A explained that they would use AI as little as possible "because it's [their] philosophy that when you want to create, it should purely come out of you." Though they were willing to use more AI suggestions "because of the time" (circumstances affecting motivation). Such use would be for when "images [and] text about the scenes that [they] have in [their] head" are "a bit vague" (mental imagery clarity). **3) Use of images and audio:** No matter the clarity, P2A mentioned that they would not use image nor audio generation as inspiration. They explained that "it's not related to the level of control". It could "interrupt [images they are] trying to visualize in [their] head". "If you look at an image, it somehow gives you some box that your creative process should be defined by this particular image." For audio, they explained that it is partly because it never "comes to [their] mind" (different mental imagery types). Though P2A mentioned that they would use generated images "like bookmarks" to visualize the structure of stories with "30 or 30 nodes plus" in an event node graph because they "wouldn't have time to read all nodes" (circumstances affecting motivation and use case for a higher level of abstraction). Their choice of using images over text is because, "for specific scenes, images come to [their] mind first" (mental imagery types). **4) Graph versus continuous text:** P2A's usual use of event node graphs was also described as depending on "the efforts of creating a graph". Specifically, while P2A mentioned usually using event node graphs similar to "Freytag's [Pyramid]" to clarify the structure of their story (mental imagery clarity), they preferred recording initial ideas in continuous text "like Word" when they "don't think it's necessary to use some nodes" (motivation).

7.2.2 Variations Between Writing Processes. **1) Chatbot versus non-chatbot (Edit Text) interfaces:** 10 other participants also preferred chatbot and non-chatbot interfaces based on their storytelling approaches. For instance, P7B preferred "conversing with the chatbot" impersonating their characters to "get inspiration about the plot [and] details about specific scenes" because they "like to use characters to construct the plot" (character-driven storytelling approach). While 4 (including P2A) out of 7 participants who adopted a more plot-driven approach mentioned a non-chatbot interface as more suitable (e.g., "Text Edit is more focused. ChatBot is more divergent." P1A), 2 participants preferred using a chatbot for practical reasons (motivation), due to greater familiarity with the interface (P5A) and/or preferred use of the conversation history to track information (P4A). No matter the approach, similar to P2A, 9 participants preferred conversing with a chatbot impersonation of their characters to write related details (level of abstraction). An exception is P5A. They explained that, if a character talks to the author, "it's kind of weird" no matter the situation (storytelling approach). **2) AI contribution:** The other participants expressed willingness to use AI if it follows their intent, does not require much prompt engineering, and/or will not lead to others doubting the quality of the work (e.g., "controversy related to authorship" P7B; motivation) for inspiration or exploration when their thoughts are described as vague and/or for validation when their thoughts are described as clear (mental imagery clarity). **3) Use of images and audio:** Similar to P2A, 6 participants who did not think that audio generation could inspire them justified this with the absence of audio in their mind (mental imagery types). Six participants explained their use of audio generation based on practical reasons (motivation), such as visualization for potential multimedia creative writing works (Section 7.1.3) and/or its potential to complement (e.g., "enrich" P1A) their thoughts (mental imagery). For image generation, 2 participants found it distracting or complementing to images in their head depending on the stages of their writing process (same mental imagery type). For instance, P3A found that images would "break the flow" of "images about story scenes" in their head when planning/translating but could "support imagination" of such images when reviewing. Similarly, the same participant who only has thoughts in text can find generated images distracting or complementing (different mental imagery types). While P5A considered images generally distracting "because [their] thoughts are just text", they admitted that they would use image generation "to write about character descriptions" since "an image could show characteristics [they] haven't thought about". Similar to P2A, 6 participants described or were observed using images as "bookmarks" (Figure 9 caption 1). **4) Graph versus continuous text:** Nine other participants, from those who usually use graphs to those who usually only write the story in continuous text, agreed that their use of an event node graph depends on necessity (motivation), for organizing their thoughts and/or for reviewing the logic of the story, and the amount of immersion they require. For instance, P6A mentioned using "a graph [can help] arrange thoughts", but when they "write something in a detailed way, [it] can break the flow" (different levels of abstraction). The use of graph editing customization features can also depend on circumstances (motivation), the audience (e.g., "I would arrange [the layout of the graph] if I have to show it to someone." P3A) and the length of the story. For the latter, 7 participants who chose node/link colors and shapes "randomly" for shorter stories would adopt a "discipline" for longer ones (quoting P5B).

7.3 Potential Factors Influencing Preferences for Information Presentation in an Event Node Graph

Participants described different preferences for the customization of node/link colors and shapes and their composition with explanations falling under five potential factors: the clarity of the text content (Section 7.3.1), other visual contrast (Section 7.3.2), the story length (Section 7.3.3), association (Section 7.3.4), and the level of abstraction (Section 7.3.5).

7.3.1 Clarity of the Text Content. 12 participants mentioned preferring the text layout in rectangular node shapes for writing event descriptions (e.g., "it's easier to read compared to other shapes" P3B), 7 preferring only using rectangular shapes for any type of information (e.g., "more shapes will clutter the board" P3A), and 3 preferring circular shapes as readable additions to differentiate specific events (Figure 9 caption 3). Similarly, for colors, some mentioned preferring ones that ensure readability (Figure 9 caption 2).

7.3.2 Other Visual Contrast. Apart from the contrast with the text content (Section 7.3.1), 11 participants described their preferences as based on contrast visually between the plot events, usually main ones, and other types of information for link shapes (e.g., "thicker" solid lines "for main plot" and "thinner" dashed lines "for secondary plot" because it is "visually prominent" P5A and Figure 9 caption 4), between colors (e.g., "something that's contrasting to the main color of the [story] section" P2A), and between node shapes (e.g., shapes for other information "visually different from story shapes" P5B). Similarly, P4A preferred colors that are visually less contrasting for what they deem to be sub-types of the same type of information. For longer stories, they envisioned using "different colors from the same color palette to represent different story events, [...] darker shades for more important events."

7.3.3 Story Length. Nine participants specified more colors and/or shapes for longer stories to categorize a greater diversity of information (e.g., event excitement levels, event importance, types of endings, alternative branches, story or character background, different characters' versions of the plot, and AI prompts if applicable), with 9 preferring using more *node colors*, 7 *link shapes* (solid and dashed), 5 *link colors*, and 3 *node shapes*.

7.3.4 Association With Specific Experiences/General Meanings. Nine participants mentioned preferring colors and/or shapes based on specific connections (e.g., red for "important" events because it "reminds [them] of fire" and "danger" P1B, red and "spiky" shapes for the same use because in "cartoons, when people get angry, they turn red [with] spiky shapes" P4A, and no "diamond" for an event node because it reminds them of "some kind of anchor point" P5B) or some general meanings they could not specify the origin of (e.g., "I can't exactly remember how, but I have this idea that, for solid, it's like the important thing; for broken lines, it's not so important" P2A

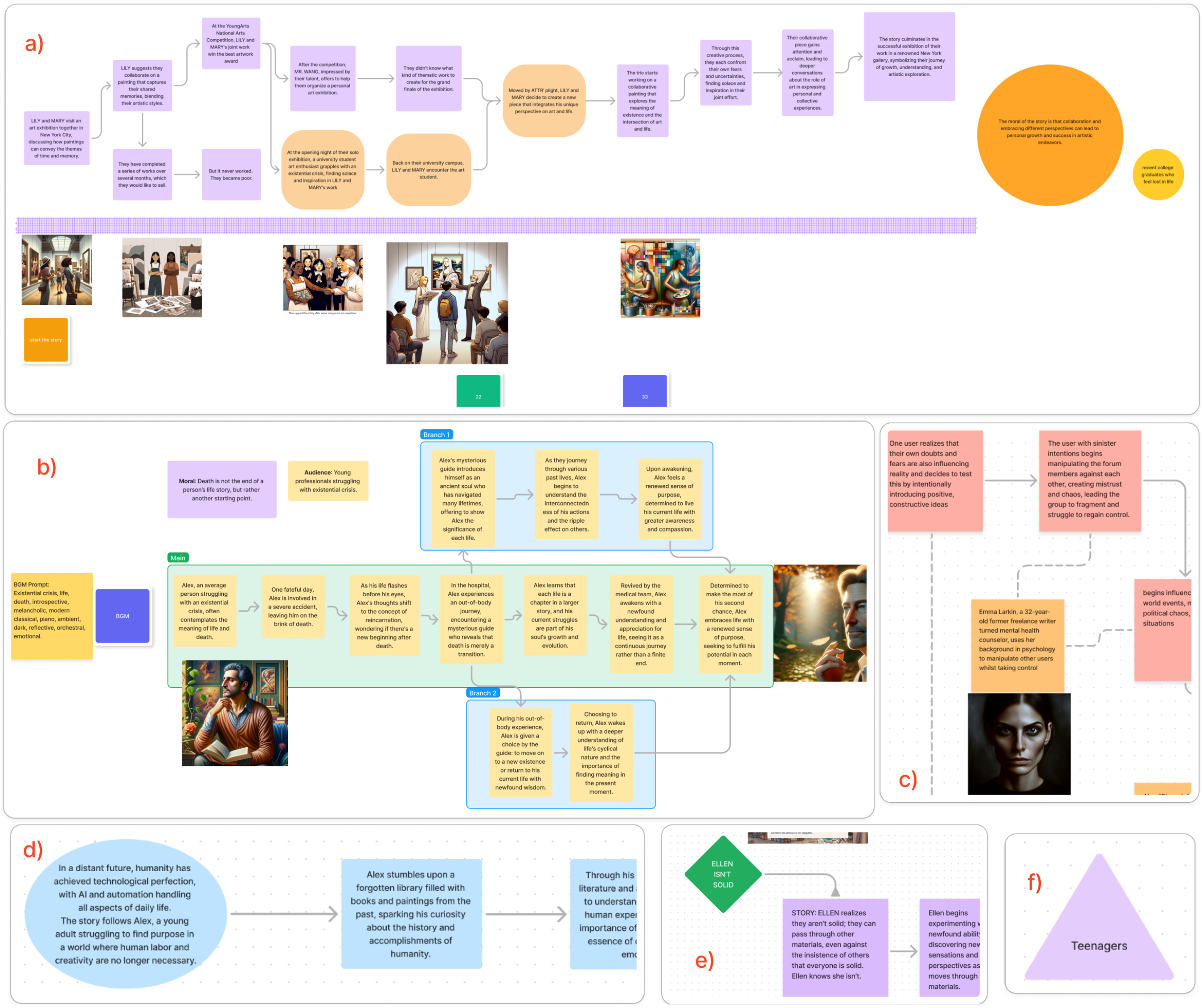

Figure 9: Screenshots of participant responses showing varied preferences for information presentation. 1) "Bookmarks": participants used images as both sources of inspiration and markers in graphs that look more like "timeline[s]" (P7B in a) or less (P2A in b), suggesting use cases for different levels of abstraction (Section 7.2). 2) Colors for readability: participants preferred different colors for nodes (e.g., "pale enough so you can see the text" P5A and "quite similar to the quality of the paper, like yellow or white" P6A) and/or node group backgrounds (e.g., "some color that's easy to look at, maybe green or blue" (b) P2A). 3) Node shapes: while some preferred "rectangular" shapes only for events (b), others also used more circular shapes (e.g., alternative versions (a) and "setting node" P7A (d)). While participants agreed to use other shapes less for event nodes, they differed on using them for information they considered "important" (e.g., triangle for audience (f) for P1B) or "less important" (e.g., diamond for title (e) for P5B). 4) Line shapes: some preferred dashed lines for relationships other than the main story's progression (e.g., details about a character (c) for P5B) for visual contrast ("It's just not as solid. So it's kind of like there, but it could be connected to lots of things." P5B).

or using black for "bad endings" because "black often represents bad things have happened" P7A).

7.3.5 Level of Abstraction. Seven participants mentioned using nodes or compositions, groups of nodes with text, image, and/or music, to visualize content of different perceived levels of importance based on "sizes", smaller for less importance (i.e., "things that are not the story itself" P7A or "details" for others). Three specified this to be visually clearer than changing node shapes.

7.4 Creativity Support Index Results

The average CSI score for FigJam/StoryNode (81.2; range: 65.3-100.0; standard deviation: 8.8) is larger than that of Twine (49.5; range: 22.7-83.0; standard deviation: 22.6; all scores in Figure 10). As our data did not follow a normal distribution (by the Shapiro-Wilk test) and our sample size is low, we conducted the *Mann-Whitney U Test* to evaluate differences between overall CSI scores and all factor scores weighted with rankings (7 statistical tests). To account for the type I error, for each statistical test, we used a Benjamin-Hochberg adjusted significance level of 0.0143. Results suggest significant difference between the overall CSI scores ($z = -3.56095$, $p = 0.00038$) and Expressiveness scores ($z = -2.5501$, $p = 0.01078$), suggesting that participants found FigJam/StoryNode to be respectively more supportive of their creative process overall and of their creative expression. Quantitatively, Expressiveness being ranked the most important factor (Figure 10) could have led to the difference in overall CSI scores. Qualitatively, the significant differences can be associated with interview responses where all participants justified their preference for FigJam/StoryNode (overall CSI) with its AI features' capabilities in helping them be "more creative" with their written expression (quoting P1A) and in supporting diverse personalized use cases for written expression (e.g., for chatbot and non-chatbot in Section 7.2; Expressiveness). Less significant differences for the other factors can be associated with more nuanced views in the qualitative data. For instance, depending on the stages, participants can find FigJam/StoryNode's breadth of features more engaging or more distracting (Section 7.1.4; Immersion). Similarly, participants found the breadth of features more supportive of their exploration and tracking of ideas (e.g., respectively through different sources of inspiration and graph versus continuous text formats; Section 7.2) and mention its potential in increasing productivity (Section 7.1.2), but they considered prompt engineering a challenge (Section 7.1.4; Exploration and Results Worth Effort). This can be further supported by the fact that the only participant who rated Twine higher (77.0 for Twine and 76.3 for FigJam/StoryNode) has a lower weighted score for "Results Worth Effort" only, ranked as the most important factor for them. This score can be associated with their reported familiarity with generative AI in the interview. The mentioned challenges for certain stages could have led to a more nuanced assessment of overall enjoyment across entire writing processes (Enjoyment). While all recognized the potential of sharing stories with human reviewers through event node graphs (Section 7.1.2), 9 participants found collaboration during the task unnecessary, giving the same Collaboration score for Twine and FigJam/StoryNode (Collaboration).

7.5 External Evaluation

Participants created 28 pairs of Twine-FigJam/StoryNode task responses (i.e., each pair with the same moral, audience, and author; average of 193 words per main story outline) diverse across morals and audience life experience and reading preferences (Table 8). Given the potential amount of resources required to find enough evaluators from each audience group, we recruited 19 creative writers of diverse cultural and creative writing thus life and literary experiences (Table 7) as external evaluators instead. We sent each evaluator an online multiple-choice questionnaire (Qualtrics) made of two parts: response quality evaluation (Section 7.5.1) and information presentation preferences (Section 7.5.2). Evaluators are each offered a compensation of about 14 USD.

7.5.1 Response Quality Evaluation. For each Twine-FigJam/StoryNode response pair (anonymized, order randomized, and fixed for readability), we asked several questions (Table 5) to understand 1) how well the moral is conveyed overall to the chosen audience and 2) how much LLM cultural biases might be related to evaluation on the overall quality and different aspects of logic (i.e., "Pacing", "Ending", and "Logical Path" in Table 5) given divided opinions on cultural differences (Section 7.1.1). For each question, the evaluator could choose "similar", the Twine outline, or the FigJam/StoryNode outline ('anonymized' as "A" and "B" to diminish biases). For 1), 14 evaluators chose FigJam/StoryNode for more comparisons, and 5, Twine. For 2), as we found no close reference, for cultural biases, we leveraged the Euclidean distances between different locations and GPT-4o in the Inglehart–Welzel World Cultural Map, a commonly used mapping of cultural values through two dimensions, traditional versus secular and survival versus self-expression [141]. We calculated Pearson correlation coefficients between the FigJam/StoryNode scores (number of times FigJam/StoryNode was chosen $+0.5\times$ number of times "similar" was chosen) of the evaluation scores and aggregate cultural distances between GPT-4o and both evaluators and authors. As seen in Table 5, we only found weak relationships, suggesting little correlation between our measures of cultural biases, each evaluator's evaluation, and the evaluated quality (overall and logic) of each author's works. Pearson correlation coefficients between the FigJam/StoryNode score for each pair for the overall evaluation and for plot logic aspects suggest moderate positive correlation ($r(26) = 0.611$ for "Pacing", $r(26) = 0.6433$ for "Ending", and $r(26) = 0.6392$ for "Logical Path"), aligning with views on the relevance of logic for conveying a moral. More evaluators chose FigJam/StoryNode for all questions on logic (13 for "Pacing", 15 for "Ending", and 15 for "Logical Path").

7.5.2 Information Presentation Preferences. Inspired by authors' diversity in information presentation preferences (Section 7.3), we also asked evaluators about such preferences (Table 11). Findings build on Section 7.3, suggesting visual cues within the same category (e.g., color) can have opposite effects on the same viewer, by being helpful if they reflect their preferences and distracting if not. We found no significant pattern between evaluators' and authors' collected demographic information and preferences (i.e., participants with the same characteristic having the same preference).

8 Discussion

Our findings add to the literature with three main novelties: on the use of LLM creative support for conveying a moral, on the combination of previously separately studied features for supporting individualized story writing processes in general, and on thinking patterns behind individual preferences for an event node graph's appearance.

Firstly, we present findings on how an LLM creative support tool could help writers fulfill an essential purpose of stories: conveying a moral (Section 7.1.1 and [14, 59, 122, 147]), a previously requested

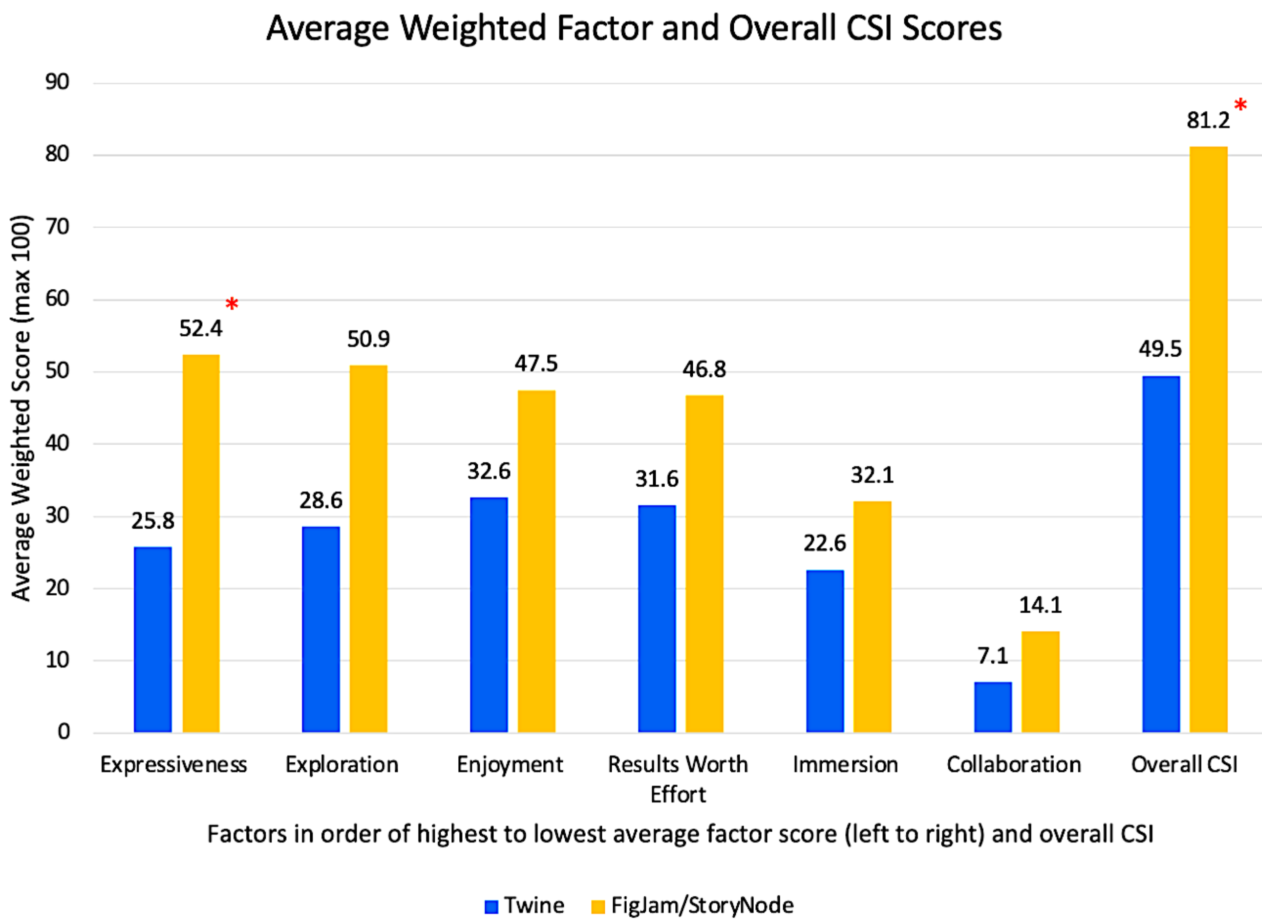

Figure 10: Average weighted factor and overall CSI scores for Twine and FigJam/StoryNode. From left to right in the figure, the factors are shown in order of highest to lowest average factor scores, used for ranking. An asterisk ("*") next to a pair of scores indicates significant difference according to the *Mann-Whitney U Test* after correction.

	Average (E)	Median (E)	Average (A)	Average (AW)	Median (A)	Median (AW)
Overall	-0.0519	0.0035	0.2158	-0.0577	0.1907	-0.0276
Pacing	-0.0154	0.2771	0.0401	-0.2847	0.1309	-0.0918
Ending	-0.074	0.0413	0.2761	-0.0959	0.3481	0.0454
Logical Path	0.0984	0.1616	0.2189	-0.0951	0.2777	0.0471

Table 5: Pearson correlation coefficients between response evaluation results and our measures of cultural biases, the averages and medians of Euclidean distances [141] between GPT-4o and locations reported by evaluators ("E") and authors, with unavailable data excluded (i.e., Cuba and Israel). For authors, we calculated both unweighted values ("A") and weighted values ("AW"), distances of locations the author specified being most influenced by only when applicable. Response evaluation results include FigJam/StoryNode scores for the overall expression ("Which OUTLINE would make a story that better conveys the MORAL to the defined AUDIENCE?") and different aspects relevant to plot logic [27] ("Which OUTLINE's story unfolds at a speed that feels more appropriate and balanced?" for "Pacing", "Which OUTLINE has a more natural and earned ending as opposed to arbitrary or abrupt?" for "Ending", and "Which OUTLINE's events follows a more logical path?" for "Logical Path"). For authors, we used the sum of FigJam/StoryNode scores from both response pairs.

but still unanswered support need [10]. To prior LLM work on story generation based on a moral, we add insights on the complementary roles of AI and humans [21, 64], with AIs suggesting general views of a greater variety of audience groups and humans adding "individualized" nuances (Section 7.1.4 and [51, 105]). Such collaboration could address recurrent concerns about social biases in humans'

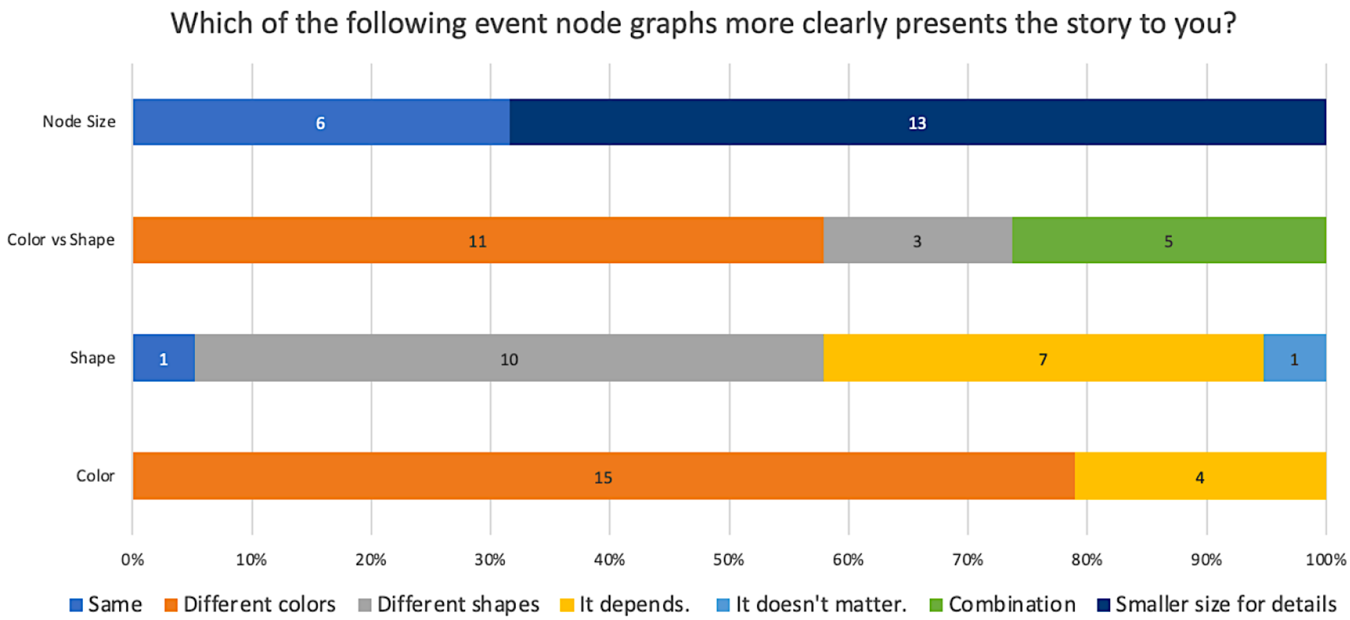

Figure 11: Results of information presentation questions answered by the external evaluators, a series of comparisons for "Which of the following event node graphs more clearly presents the story to you?" (more in Section B.3). Starting from the bottom, the choices for "Color" are "all the same node color and shape" (n=0), "nodes with the same shape but different colors to represent different types of information" (n=15), "It depends. If the nodes are colored based on how I would color them, they can help me understand the story more quickly. If the nodes are colored differently, even if there are instructions, it takes more effort to develop understanding. So, it can be distracting." (n=4), and "It doesn't matter. When I look at a node graph, colors as visual cues are irrelevant to me." (n=0). Similarly, for "Shape", the choices are "all the same..." (n=1), "nodes with the same color but different shapes..." (n=10), "It depends..." (n=7), and "It doesn't matter..." (n=1). For "Color vs Shapes", the choices are "nodes with the same shape but different colors..." (n=11), "nodes with the same color but different shapes..." (n=3), and "A combination (using both colors and shapes) would be the clearest." (n=5). For "Node Size", the choices are "same size" (n=6) and "smaller size for the node containing details about the event" (n=13).

and AIs' writing [18, 27, 61, 141]. For human-AI collaboration, we present a tool design whose capabilities for supporting written expression, reflection on story logic, and ultimately expression of a moral to a specific audience are grounded in theory (Section 2) and supported by empirical data. Specifically, qualitative and quantitative data suggest that not only did authors find our tool supportive of both written expression (e.g., Expressiveness explained in Section 7.4) and reflection on story logic (e.g., visualization through graph editing in Sections 7.1.1 and 7.2), external evaluators also found that outputs created with our tool can better convey the moral to different audience groups and preferred their story logic (Section 7.5.1). Our design could be a starting point for research leveraging either current or future technologies [102, 103]. Given the weight both industry and academia put on morals' potential impact [14, 95, 121, 144, 147], such research could focus on context-specific criteria (e.g., a specific class' academic rubric). To prior work supporting the potential of LLM-powered node graph editing tools for storytelling in general [110, 155], we add nuances on the need for graph visualization, which can depend on the overall goal (e.g., conveying a moral in Section 7.1.4) and sub-goals at specific writing stages (e.g., motivation in Section 7.2.2.4), in line with views on needs shifting based on higher level goals and sub-goals [18, 47, 51]. For prior LLM writing support works, which studied graph editing and impersonation separately [18, 77, 110, 155], our findings (Section 7.1.1) also suggest that an interplay between impersonation and graph editing (e.g., Figure 8) could lead to preferred writing experience (Section 7.4) and even output (Section 7.5.1) beyond story writing around a moral (Section 7.1.2). Though, building upon the LLM impersonation research [18], our findings further suggest that prompt engineering for audience personas depends on writing goals with nuances specific to conveying a story's moral across cultures (Section 7.1.1), such as the inclusion of fiction story preferences (e.g., "interested in magical creatures") and the absence of culture specification (e.g., "from China"). While more specific prompts can diminish model biases [141], writers might have difficulties creating them, as seen with lower "Results Worth Effort" scores in Section 7.4 and participant feedback (Section 7.1.4 and [18]). Future research could study differences in audience persona prompts between writing goals, such as conveying a moral versus expressing oneself in a personal journal, to inform (automated) prompt engineering. Further research can also focus on cultural nuances, possibly on how values that are more differently prioritized [54, 153] can introduce nuances among story writers' personas and morals. We discuss how our tool and study design could be leveraged for such research in Section 8.1.1. To explore how a tool could accommodate shifting feature needs (e.g., between impersonation and graph editing), research could study their interplay.

Secondly, we present findings on how an interplay between previously separately studied features - impersonation through chatbot and non-chatbot interfaces, graph editing, and image and audio generation - could align with shifting needs across writing processes [47, 51] for story writing in general (given overlapping needs; Sections 2 and 7.1.2). Future works could further investigate how such interplay could improve writing experience (discussed in Section 8.1.2), mitigate the impact of biases through interface design (discussed in Section 8.1.1), or improve writing output. Though features' usage patterns can be unique among writers (e.g., Section 7.2 and [21, 129]) and change with AI models' capabilities (e.g., older non-LLM versus recent LLM [108]). For greater comparability, we connect such usage patterns to higher-level factors (Section 6.3), which separately correspond to usage patterns found in prior AI writing support works: mental imagery and storytelling approaches for text and image generation and LLM character impersonation [108] for example, motivation for audience impersonation [18] and integration of AI content based on circumstances [129], and levels of abstraction for graph visualization of relationships and AI generation for more specific story elements [110, 155]. Our findings on the interplay between factors also complement cognitive process research on the diversity of brain functions required for moral of the story appreciation [94], motivation as a balance between costs and benefits, both dependent on the audience, during writing [60], creative support design to prevent overloading users' working memory by allowing them to "offload" certain information in their environment or tools (i.e., Distributed Cognition [41]) with preferences for continuous text over graph formats and conversation history to track information for instance, variations across levels of abstraction [47], and variations among mental imagery types, clarity, and sources of inspiration during writing (Section 2.2). Additionally, our findings suggest varying amounts of influence from different factors between writers. This can be illustrated by P5A's consistency in the use of the same LLM chatbot interface due to practicality (motivation) and their storytelling approach (Section 7.2.2) versus P2A's changes between LLM interfaces due to different levels of abstraction despite their storytelling approach (Section 7.2.1). Future research could further validate our factors with other user groups (e.g., culturally different), different weights for different writer profiles, and their relationships to preferences for a graph, a continuous text outline, or the full text across evaluation criteria (e.g., plot logic versus quality of written expression in Section 7.1.4). As some interplays might only be observable through our breadth of features (e.g., interplays between all features corresponding to interplays between all four factors in Section 7.2.1), for more comprehensive insights, research can include our breadth of features, ideally in a single tool [70, 112].

Thirdly, we add findings on thinking patterns behind individual preferences for an event node graph's node/link appearance to prior work on graph visualization of creative writing stories. Specifically, our findings suggest that, if presented the opportunity, authors and reviewers might prefer individualized combinations of node/link colors and shapes and spatial arrangement (e.g., timelines versus freer arrangement in Figure 9) to differentiate between information (e.g., in contrast to only line colors for line graphs [140] or spatial arrangement for node graphs with no such node/link customization [110, 155]). Though such combinations can have opposite effects, be distracting, when they do not reflect the viewer's preferences (Section 7.5.2). Further research on accommodating such visualization preferences could thus improve not only the writing experience (for support tools) but also communication (for visualization in general). While patterns among specific visualization preferences (e.g., preference for colors over shapes) seem absent (Section 7.5.2), such preferences can reflect potential higher-level factors (Section 7.3) grounded in theory and empirical data (Section 6.3). Thus, to works that did mention specific preferences for node/link colors or shapes (Section 2.4), we add such factors, which could improve comparability among individualized visualization preferences. In particular, such factors are in line with findings on cognitive processes behind visualization, highlighting the need for further research on supporting individual differences [79] in spatial ability (i.e., understanding and memorization of spatial relations among objects) possibly through node/link sizes based on the importance of the content (Section 7.3.5), in associative memory (i.e., the ability to remember a relationship between two seemingly unrelated objects as seen in Section 7.3.4, which can be influenced by cultural experience [137]) possibly through a culturally diverse writer group, and in perceptual speed (i.e., rate at which one identifies figures or symbols) possibly through the clarity of the text and between nodes/links (Sections 7.3.1 and 7.3.2) for stories of varying complexities (Section 7.3.3).

8.1 Design Implications

8.1.1 AI Support for Story Writing Around a Moral and Cultural Bias Study. In line with the universality of some morals, such as those promoting care [61, 122, 153] (e.g., "mutual understanding" in Table 8), our findings suggest no significant instance of cultural bias (Sections 7.1 and 7.5.1), but concerns and the shared goal of conveying morals across cultures (Section 7.1.1 and [5, 84, 138, 159]) warrant further research on mitigating model biases' impact. This can be done through 1) fine-tuning, which can require extracting causal inferences and meanings [61, 153], 2) prompt engineering, which can require understanding nuances between languages and output structures [141], and 3) mitigating homogenization [3] through an interface design encouraging reflection on creative choices [72], which requires connecting usage patterns to reflection on the moral conveyed for instance [52]. All these require understanding of cultural nuances in logic and written expression.

To do so, future research can leverage our tool's features together or separately, through the lens of our potential higher-level factors (Sections 7.2 and 7.3) for possibly greater comparability (Section 8). By asking writers to express their story through graph editing, researchers can analyze nuances in their logic understanding. For instance, the writer's use of a color they associate with good or bad for an ending (e.g., Section 7.3.4) could reflect cross-cultural differences in causal inferences (e.g., Section 7.1.1 and [141, 150]). Given cross-cultural differences in AI content integration for the same interface [3] and different usage patterns for different interfaces (e.g., more conversational for chatbot in Section 7.2.2 and [112]), research could discover cultural nuances through writers' interaction with different interfaces across levels of abstraction (Section 7.2), for logic (outline from higher level) and written expression (story scenes from lower level). For instance, for logic, the integration of plot ideas obtained through chatbot conversation with a character impersonation (e.g., P7B in Section 7.2.2.1) versus through a command-like request in a non-chatbot (e.g., P2A in Section 7.2.1.1) could reflect connections between reliance on AI, storytelling approaches, and cross-cultural differences in causal inferences (i.e., personal traits versus contextual factors in Section 2.2). For written expression, integration of character conversation into a story scene (e.g., P2A in Section 7.2.1.1) could reflect reliance more due to the AI content's perceived cultural closeness (Section 7.1.1 and [3]) with the characters talking than storytelling approaches. Such nuances could inform different prompt recommendations, through different keywords between audience and character personas (e.g., more culturally specific for characters [108]) for instance, and strategies for encouraging reflection, such as through AI-generated questions [52, 154] (e.g., impersonated character asking about the plot versus written expression). Similarly, research could study how cultural biases in non-textual cues (e.g., images [160]) influence written expression as direct sources of inspiration (e.g., P5A in Section 7.2.2.3) and understanding of the logic as markers (e.g., Figure 9.1). Given different dynamics [18, 51], research could also explore collaboration between humans (e.g., authors and reviewers) of different cultures, possibly comparing with AI collaboration only for biases toward the author's identity (Section 7.2.2.2 and [83, 161]). Our plugin, for a platform supporting real-time collaboration, can be used.

For study design, to draw more connections, researchers could collect additional cultural background information, such as the weighted influence of different cultures on the writer (Section 7.5.1), other cultural factors (e.g., ideological beliefs [54]), and technology use since users can be influenced by values spread through cyberspace (e.g., social media and online stories [13, 156]). Future research can also adapt our tasks for more specific findings (e.g., audience of a specific culture), drawing comparisons with our quantitative findings (Sections 7.4 and 7.5.1).

As cultural biases concerns can be for writing in general (Section 8), findings can be relevant as well.

8.1.2 Personalized AI Creative Writing Support.

As writers might prefer a system that selectively displays specific features based on shifting needs across story writing processes (e.g., Section 7.1.4 and [108]), future research could seek to connect our potential factors affecting usage patterns (Section 7.2) to writer profiles, customizable input fields for generative AI (e.g., art style selection for image generation [108] and attribute fields for personas [18]) to reduce prompt engineering difficulties (e.g., Section 7.4), and prompt suggestion preferences, complementing works on automated prompt optimization (e.g., [76, 119]).

8.1.3 Personalized Visualization.

As personalization to individualized visualization needs can affect experience (Section 8), future research could explore a 'translator' feature that bidirectionally converts a given graph to another or continuous text based on user profiles defined by factors similar to ours (Section 7.3), validate these factors with a culturally diverse group given nuances [16] through metrics for visualization abilities [79], explore how different parts of the continuous text story correspond to graph components (e.g., the protagonist's fortune [36] and story relationships [155]), and explore nuances between writing tasks (e.g., detective versus romance stories; Section 7.1.4).

8.2 Limitations

By focusing on a balance between exploration and resource availability, our work thus leads to opportunities to support directions found with more empirical evidence. This can be for specific AI models, writer groups (e.g., specific cultural backgrounds or levels of familiarity with AI), stories of varying lengths, full stories (e.g., through longitudinal studies as longer stories could take months to write [33, 87]), audience groups (e.g., finding a statistically significant number of readers for each writer participant's chosen audience), and in-the-wild settings [116].

9 Ethical Considerations

While the institutional review board has approved our research, ethical concerns might still arise. We describe how we addressed them for future reference. First, generative AI content could disturb some participants [69]. We used commercial generative AI with filters [98, 99, 134], informed every potential participant about the risks, and mentioned the possibility to withdraw anytime. Second, human authors' writing samples could disturb some evaluators and raise concerns about the authorship rights. For the former, we required authors to exclude "explicit sexual and/or strong, disturbing violent content" and manually verified all samples. For authorship, we informed authors of AI platforms' terms of use and evaluators, of the authors' rights over their works. Third, for participant privacy, we only shared data and works for which we have received consent to share. No further concern was raised.

10 Conclusion

We studied how a single tool can support reflection on a moral alongside other story writing needs. Through a formative study (N=12), a user study (N=14), and external evaluation (N=19), we designed, implemented, then studied StoryNode, a prototype plugin for FigJam. FigJam/StoryNode supports visualization of the story structure through customizable node graph editing, LLM impersonation (chatbot and non-chatbot interfaces), and image and audio generative AI features. Our findings support such a tool's potential for story writing in general. They also include potential factors affecting the interplay between features, which could inform personalized creative writing support design and story visualization.

Acknowledgments

This research was partly supported by a grant from the Guangzhou Municipal Nansha District Science and Technology Bureau under Contract No.2022ZD01. We thank all reviewers and participants for their time. Figures 1, 2, 5, 6, 7, 8, 9, and 12 are made with FigJam [46]. Figures 1, 6, 8, and 9 contain output created with OpenAI's models [100]. Figure 1 contains output generated through Suno [136]. We are not funded by the mentioned platforms.

References

[1] Hussein Karam Hussein Abd El-Sattar. 2010. A new plot/character-based interactive system for story-based virtual reality applications. *International Journal of Image and Graphics* 10, 01 (2010), 113–133. https://doi.org/10.1142/S021946781000369X

[2] C.E. Acosta, C.A. Collazos, L.A. Guerrero, J.A. Pino, H.A. Neyem, and Moteletm O. 2004. StoryMapper: a multimedia tool to externalize knowledge. In *XXIV International Conference of the Chilean Computer Science Society*. 133–140. https://doi.org/10.1109/QEST.2004.21

[3] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2024. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. (2024). https://doi.org/10.48550/arXiv.2409.11360 arXiv:2409.11360

[4] Taisuke Akimoto. 2016. Prototyping a Narrative Structure Interface as a Basis for Human-Computer Co-creation of Narratives. *International Journal of Knowledge Engineering* 2, 4 (2016), 158–164. https://doi.org/10.18178/ijke.2016.2.4.072

[5] Ali Al-Jafar and Cary A Buzzelli. 2004. The art of storytelling for cross cultural understanding. *Int. J. Early Child.* 36, 1 (June 2004), 35–48. https://doi.org/10.1007/BF03165939

[6] Latifa Al-Naimi and Mirela Alistar. 2024. Understanding Cultural and Religious Values Relating to Awareness of Women's Intimate Health among Arab Muslims. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 690, 18 pages. https://doi.org/10.1145/3613904.3642207

[7] Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.* 54, 5, Article 103 (may 2021), 38 pages. https://doi.org/10.1145/3453156

[8] Linda Anderson, Mary Hammond, Sara Haslam, W.N. Herbert, Derek Neale, and W.R. Owens. 2006. *Creative Writing: A Workbook with Readings*. Routledge.

[9] Anthropic. 2024. *Introducing the next generation of Claude.* Retrieved August 21, 2024 from https://www.anthropic.com/news/claude-3-family

[10] Victor Nikhil Antony and Chien-Ming Huang. 2024. ID.8: Co-Creating Visual Stories with Generative AI. *ACM Trans. Interact. Intell. Syst.* 14, 3, Article 20 (aug 2024), 29 pages. https://doi.org/10.1145/3672277

[11] Atsushi Ashida and Tomoko Kojiri. 2019. Plot-creation support with plot-construction model for writing novels. *Journal of Information and Telecommunication* 3, 1 (2019), 57–73. https://doi.org/10.1080/24751839.2018.1531232

[12] Claus Atzenbeck, Sam Brooker, and Daniel Roßner. 2023. Storytelling Machines. In *Proceedings of the 6th Workshop on Human Factors in Hypertext* (Rome, Italy) *(HUMAN '23)*. Association for Computing Machinery, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3603607.3613481

[13] Magdalena Baga. 2023. Digital Novels: A Recycled Advertisement about the Old Social Construction of Women's Identity. *TRANS-KATA: Journal of Language, Literature, Culture and Education* 3, 2 (May 2023), 93–103. https://doi.org/10.54923/jllce.v3i2.46

[14] Michael J. Baker, Gwen Pallarès, Talli Cedar, Noa Brandel, Lucas Bietti, Baruch Schwarz, and Françoise Détienne. 2023. Understanding the moral of the story: Collaborative interpretation of visual narratives. *Learning, Culture and Social Interaction* 39 (2023), 100700. https://doi.org/10.1016/j.lcsi.2023.100700

[15] Michael Balas, Jordan Joseph Wadden, Philip C Hébert, Eric Mathison, Marika D Warren, Victoria Seavilleklein, Daniel Wyzynski, Alison Callahan, Sean A Crawford, Parnian Arjmand, and Edsel B Ing. 2024. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *Journal of Medical Ethics* 50, 2 (2024), 90–96. https://doi.org/10.1136/jme-2023-109549

[16] Lyn Bartram, Abhisekh Patra, and Maureen Stone. 2017. Affective Color in Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1364–1374. https://doi.org/10.1145/3025453.3026041

[17] Zekerya Batur. 2018. The Analysis of the Level of Students' Perception of the Messages in Fictive Texts in Fictional Context. *Reading Improvement* 55, 2 (2018), 69–78.

[18] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1049, 18 pages. https://doi.org/10.1145/3613904.3642406

[19] William J. Bennett. 1993. *The Book of Virtues* (1st ed.). Simon & Schuster, New York.

[20] Lina Bentahila, Roger Fontaine, and Valérie Pennequin. 2021. Universality and cultural diversity in moral reasoning and judgment. *Front. Psychol.* 12 (Dec. 2021), 764360. https://doi.org/10.3389/fpsyg.2021.764360

[21] Oloff C. Biermann, Ning F. Ma, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1209–1227. https://doi.org/10.1145/3532106.3533506

[22] Fearn Bishop, Johannes Zagermann, Ulrike Pfeil, Gemma Sanderson, Harald Reiterer, and Uta Hinrichs. 2020. Construct-A-Vis: Exploring the Free-Form Visualization Processes of Children. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 451–460. https://doi.org/10.1109/TVCG.2019.2934804

[23] Ali Borji and Mehrdad Mohammadian. 2023. Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. (June 2023). https://doi.org/10.2139/ssrn.4476855

[24] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. (2023). https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712

[25] Emma E. Buchtel, Yanjun Guan, Qin Peng, Yanjie Su, Biao Sang, Sylvia Xiaohua Chen, and Michael Harris Bond. 2015. Immorality East and West: Are Immoral Behaviors Especially Harmful, or Especially Uncivilized? *Personality and Social Psychology Bulletin* 41, 10 (2015), 1382–1394. https://doi.org/10.1177/0146167215595606

[26] Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2019. *Writing Fiction: A Guide to Narrative Craft* (10th ed.). University of Chicago Press. https://doi.org/10.7208/chicago/9780226616728.001.0001

[27] 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* (, Honolulu, HI, USA,) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 30, 34 pages. https://doi.org/10.1145/3613904.3642731

[28] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. (2024). https://doi.org/10.48550/arXiv.2310.00785 arXiv:2310.00785v3

[29] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (mar 2024), 45 pages. https://doi.org/10.1145/3641289

[30] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. 2019. Neural Storyboard Artist: Visualizing Stories with Coherent Image Sequences. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2236–2244. https://doi.org/10.1145/3343031.3350571

[31] Justin Cheng, Laewoo Kang, and Dan Cosley. 2013. Storeys: designing collaborative storytelling interfaces. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 3031–3034. https://doi.org/10.1145/2468356.2479603

[32] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.* 21, 4, Article 21 (June 2014), 25 pages. https://doi.org/10.1145/2617588

[33] Dave Chesson. 2022. *How Long Does it Take to Write a Book?* Retrieved August 21, 2024 from https://kindlepreneur.com/how-long-to-write-a-book/

[34] Pei-Yu Chi and Henry Lieberman. 2011. Intelligent assistance for conversational storytelling using story patterns. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) *(IUI '11)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/1943403.1943438

[35] Cheng-Han Chiang and Hung yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics, 15607–15631. https://doi.org/10.48550/arXiv.2305.01937

[36] 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* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. https://doi.org/10.1145/3491102.3501819

[37] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (2017), 297–298. https://doi.org/10.1080/17439760.2016.1262613

[38] Daniel Cox. 2022. *We Make How We Learn: The Role of Community in Authoring Tool Longevity.* Springer International Publishing, Cham, 65–72. https://doi.org/10.1007/978-3-031-05214-9_5

[39] 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* (, San Francisco, CA, USA,) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. https://doi.org/10.1145/3586183.3606772

[40] Susan Daniels-McGhee and Gary A. Davis. 1994. The Imagery-Creativity Connection. *The Journal of Creative Behavior* 28, 3 (1994), 151–176. https://doi.org/10.1002/j.2162-6057.1994.tb01189.x

[41] Nicholas Davis, Holger Winnemöller, Mira Dontcheva, and Ellen Yi-Luen Do. 2013. Toward a cognitive theory of creativity support. In *Proceedings of the 9th ACM Conference on Creativity & Cognition* (Sydney, Australia) *(C&C '13)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2466627.2466655

[42] Figma Developers. 2024. *Introduction*. Retrieved August 21, 2024 from https://www.figma.com/plugin-docs/

[43] Cambridge Dictionary. [n. d.]. MORAL | English meaning. In *dictionary.cambridge.org*. https://dictionary.cambridge.org/dictionary/english/moral

[44] Danica Dillion, Debanjan Mondal, Niket Tandon, and Kurt Gray. 2024. Large Language Models as Moral Experts? GPT-4o Outperforms Expert Ethicist in Providing Moral Guidance. https://doi.org/10.31234/osf.io/w7236

[45] Henrik Engstr̊om, Jenny Brusk, and Patrik Erlandsson. 2018. Prototyping Tools for Game Writers. *The Computer Games Journal* 7 (2018), 153–172. https://doi.org/10.1007/s40869-018-0062-y

[46] Figma. 2024. The Online Collaborative Whiteboard for Teams. https://www.figma.com/figjam/

[47] Linda Flower and John R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32, 4 (Dec. 1981), 365–387.

[48] Linda Flower and John R. Hayes. 1984. Images, Plans, and Prose: The Representation of Meaning in Writing. *Written Communication* 1, 1 (1984), 120–160. https://doi.org/10.1177/0741088384001001006

[49] Interactive Fiction Technology Foundation. 2024. *Twine: An open-source tool for telling interactive, nonlinear stories*. Retrieved March 28, 2024 from https://twinery.org/

[50] Eduardo C. Garrido-Merchán, José Luis Arroyo-Barrigüete, and Roberto Gozalo-Brihuela. 2023. Simulating H.P. Lovecraft horror literature with the ChatGPT large language model. (2023). https://doi.org/10.48550/arXiv.2305.03429 arXiv:2305.03429

[51] Katy Ilonka Gero, Tao Long, and Lydia B. Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages. https://doi.org/10.1145/3544548.3580782

[52] Julius Goth, Eun Ha, and James Lester. 2011. Towards a Model of Question Generation for Promoting Creativity in Novice Writers. In *2011 AAAI Fall Symposium Series*.

[53] Jonathan Gottschall. 2012. *The Storytelling Animal: How Stories Make Us Human*. Houghton Mifflin Harcourt, New York.

[54] Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, and Li Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology* 8 (2016), 125–130. https://doi.org/10.1016/j.copsyc.2015.09.007 Culture.

[55] Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101, 2 (2011), 366–385. https://doi.org/10.1037/a0021847

[56] Jonathan Haidt, Silvia Helena Koller, and Maria G Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65, 4 (1993), 613–628. https://doi.org/10.1037/0022-3514.65.4.613

[57] Hyemin Han. 2023. Potential benefits of employing large language models in research in moral education and development. *Journal of Moral Education* 0, 0 (2023), 1–16. https://doi.org/10.1080/03057240.2023.2250570

[58] Frank Allan Hansen, Karen Johanne Kortbek, and Kaj Grønbæk. 2008. Mobile Urban Drama – Setting the Stage with Location Based Technologies. In *Interactive Storytelling*, Ulrike Spierling and Nicolas Szilas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 20–31. https://doi.org/10.1007/978-3-540-89454-4_4

[59] Yuval Noah Harari. 2018. *Sapiens: A Brief History of Humankind* (reprint ed.). Harper Perennial.

[60] John R Hayes. 1996. A new framework for understanding cognition and affect in writing. In *The science of writing: Theories, methods, individual differences, and applications*, C Michael Levy and Sarah Ransdell (Eds.). Lawrence Erlbaum Associates, Inc, 1–27.

[61] David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story Morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 12998–13032. https://doi.org/10.18653/v1/2024.emnlp-main.723

[62] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (, Virtual Event, Australia,) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1260–1276. https://doi.org/10.1145/3532106.3533526

[63] Md Naimul Hoque, Bhavya Ghai, Kari Kraus, and Niklas Elmqvist. 2023. Portrayal: Leveraging NLP and Visualization for Analyzing Fictional Characters. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (, Pittsburgh, PA, USA,) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 74–94. https://doi.org/10.1145/3563657.3596000

[64] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 71 (Sept. 2023), 29 pages. https://doi.org/10.1145/3534561

[65] David Jackson and Annabel Latham. 2022. Talk to The Ghost: The Storybox methodology for faster development of storytelling chatbots. *Expert Systems with Applications* 190 (2022), 116223. https://doi.org/10.1016/j.eswa.2021.116223

[66] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 28458–28473. https://proceedings.neurips.cc/paper_files/paper/2022/file/b654d6150630a5ba5df7a55621390daf-Paper-Conference.pdf

[67] Melanie Killen and Audun Dahl. 2021. Moral Reasoning Enables Developmental and Societal Change. *Perspect. Psychol. Sci.* 16, 6 (nov 2021), 1209–1225. https://doi.org/10.1177/1745691620964076

[68] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus Gross, and Hanspeter Pfister. 2018. Visualizing Nonlinear Narratives with Story Curves. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 595–604. https://doi.org/10.1109/TVCG.2017.2744118

[69] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1046, 15 pages. https://doi.org/10.1145/3613904.3642693

[70] Tae Soo Kim, Arghya Sarkar, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. LMCanvas: Object-Oriented Interaction to Personalize Large Language Model-Powered Writing Environments. (2023). https://doi.org/10.48550/arXiv.2303.15125 arXiv:2303.15125

[71] Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. (2024). https://doi.org/10.48550/arXiv.2404.01261 arXiv:2404.01261

[72] Max Kreminski. 2024. The Dearth of the Author in AI-Supported Writing. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants* (Honolulu, HI, USA) *(In2Writing '24)*. Association for Computing Machinery, New York, NY, USA, 48–50. https://doi.org/10.1145/3690712.3690725

[73] Jorge Leandro, Sudha Rao, Michael Xu, Weijia Xu, Nebosja Jojic, Chris Brockett, and Bill Dolan. 2023. GRIM: GRaph-based Interactive narrative visualization for gaMes. (2023). https://doi.org/10.48550/arXiv.2311.09213 arXiv:2311.09213

[74] Jorge Leandro, Sudha Rao, Michael Xu, Weijia Xu, Nebojsa Jojic, Chris Brockett, and Bill Dolan. 2024. GENEVA: GENErating and Visualizing branching narratives using LLMs. In *2024 IEEE Conference on Games (CoG)*. 1–5. https://doi.org/10.1109/CoG60054.2024.10645625

[75] Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024. EMONA: Event-level Moral Opinions in News Articles. (2024). https://doi.org/10.48550/arXiv.2404.01715 arXiv:2404.01715

[76] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to Rewrite Prompts for Personalized Text Generation. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3367–3378. https://doi.org/10.1145/3589334.3645408

[77] Ge Li, Danai Vachtsevanou, Jérémy Lemée, Simon Mayer, and Jannis Strecker. 2024. Reader-aware Writing Assistance through Reader Profiles. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 344–350. https://doi.org/10.1145/3648188.3675152

[78] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[79] Zhengliang Liu, R. Jordan Crouser, and Alvitta Ottley. 2020. Survey on Individual Differences in Visualization. *Computer Graphics Forum* 39, 3 (2020), 693–712. https://doi.org/10.1111/cgf.14033

[80] Shirley Long and Elfrieda H. Hiebert. 1985. Effects of awareness and practice in mental imagery on creative writing of gifted children. In *Issues in Literacy: A research perspective (34th Yearbook of the National Reading Conference)*, J. Niles

and R. Lalik (Eds.). 381–385.

[81] JGraph Ltd. 2024. *draw.io*. Retrieved November 30, 2024 from https://www.drawio.com/

[82] Christopher J. Lynch, Erik Jensen, Madison H. Munro, Virginia Zamponi, Joseph Martinez, Kevin O'Brien, Brandon Feldhaus, Katherine Smith, Ann Marie Reinhold, and Ross Gore. 2024. GPT-4 Generated Narratives of Life Events using a Structured Narrative Prompt: A Validation Study. (2024). https://doi.org/10.48550/arXiv.2402.05435 arXiv:2402.05435

[83] Federico Magni, Jiyoung Park, and Melody Manchi Chao. 2024. Humans as creativity gatekeepers: Are we biased against AI creativity? *J. Bus. Psychol.* 39, 3 (June 2024), 643–656. https://doi.org/10.1007/s10869-023-09910-x

[84] Georgios Makridis, Athanasios Oikonomou, and Vasileios Koukos. 2024. FairyLandAI: Personalized Fairy Tales utilizing ChatGPT and DALLE-3. (2024). https://doi.org/10.48550/arXiv.2407.09467 arXiv:2407.09467

[85] Raymond A. Mar and Keith Oatley. 2008. The Function of Fiction is the Abstraction and Simulation of Social Experience. *Perspectives on Psychological Science* 3, 3 (2008), 173–192. https://doi.org/10.1111/j.1745-6924.2008.00073.x

[86] Marcel Marti, Jodok Vieli, Wojciech Witoń, Rushit Sanghrajka, Daniel Inversini, Diana Wotruba, Isabel Simo, Sasha Schriber, Mubbasir Kapadia, and Markus Gross. 2018. CARDINAL: Computer Assisted Authoring of Movie Scripts. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 509–519. https://doi.org/10.1145/3172944.3172972

[87] MasterClass. 2022. *How Long Does It Take to Write a Book?* Retrieved August 21, 2024 from https://www.masterclass.com/articles/how-long-does-it-take-to-write-a-book

[88] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. https://doi.org/10.1145/3544548.3581225

[89] Alex Mitchell and Kevin McGee. 2009. Designing hypertext tools to facilitate authoring multiple points-of-view stories. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (Torino, Italy) *(HT '09)*. Association for Computing Machinery, New York, NY, USA, 309–316. https://doi.org/10.1145/1557914.1557966

[90] John T. Murray and Anastasia Salter. 2022. *Mapping the Unmappable: Reimagining Visual Representations of Interactive Narrative*. Springer International Publishing, Cham, 171–190. https://doi.org/10.1007/978-3-031-05214-9_11

[91] Tru Narla. 2021. *Soundboard*. Retrieved August 21, 2024 from https://www.figma.com/community/widget/1028911374358427438/soundboard

[92] Darcia Narvaez. 2001. Moral Text Comprehension: Implications for education and research. *Journal of Moral Education* 30, 1 (2001), 43–54. https://doi.org/10.1080/03057240120033802

[93] Robert J. Nash. 1997. *Answering the "Virtuecrats": A Moral Conversation on Character Education.* Teachers College Press, New York.

[94] Paolo Nichelli, Jordan Grafman, Pietro Pietrini, Kimberley Clark, Kyu Young Lee, and Robert Miletich. 1995. Where the brain appreciates the moral of a story. *NeuroReport* 6, 17 (1995), 2309–2313. https://doi.org/10.1097/00001756-199511270-00010

[95] NobelPrize.org. 2024. *Facts on the Nobel Prize in Literature*. Retrieved March 28, 2024 from https://www.nobelprize.org/prizes/facts/facts-on-the-nobel-prize-in-literature/

[96] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[97] OpenAI. 2024. *DALL-E 3*. Retrieved August 21, 2024 from https://openai.com/index/dall-e-3/

[98] OpenAI. 2024. *Hello GPT-4o*. Retrieved August 21, 2024 from https://openai.com/index/hello-gpt-4o/

[99] OpenAI. 2024. *Image generation*. Retrieved November 19, 2024 from https://platform.openai.com/docs/guides/images

[100] OpenAI. 2024. *Usage policies*. Retrieved January 26, 2025 from https://openai.com/policies/usage-policies

[101] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 10 pages. https://doi.org/10.1145/3411763.3450391

[102] Antti Oulasvirta and Kasper Hornbæk. 2016. HCI Research as Problem-Solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4956–4967. https://doi.org/10.1145/2858036.2858283

[103] Antti Oulasvirta and Kasper Hornbæk. 2022. Counterfactual Thinking: What Theories Do in Design. *International Journal of Human–Computer Interaction* 38, 1 (2022), 78–92. https://doi.org/10.1080/10447318.2021.1925436

[104] Kalpesh Padia, Kaveen Herath Bandara, and Christopher G. Healey. 2019. A system for generating storyline visualizations using hierarchical task network planning. *Computers & Graphics* 78 (2019), 64–75. https://doi.org/10.1016/j.cag.2018.11.004

[105] Emmanouil Papagiannidis, Ida Merete Enholm, Chirstian Dremel, Patrick Mikalef, and John Krogstie. 2023. Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes. *Information Systems Frontiers* 25, 1 (Feb. 2023), 123–141. https://doi.org/10.1007/s10796-022-10251-y

[106] Joel Pearson, Thomas Naselaris, Emily A Holmes, and Stephen M Kosslyn. 2015. Mental imagery: Functional mechanisms and clinical applications. *Trends Cogn. Sci.* 19, 10 (Oct. 2015), 590–602. https://doi.org/10.1016/j.tics.2015.08.003

[107] Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (Almost) Dead. (2023). https://doi.org/10.48550/arXiv.2309.09558 arXiv:2309.09558

[108] 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642105

[109] Qualtrics. 2024. *Qualtrics XM: The Leading Experience Management Software*. Retrieved August 27, 2024 from https://www.qualtrics.com

[110] Ahmed Y. Radwan, Khaled M. Alasmari, Omar A. Abdulbagi, and Emad A. Alghamdi. 2024. SARD: A Human-AI Collaborative Story Generation. (2024). https://doi.org/10.48550/arXiv.2403.01575 arXiv:2403.01575

[111] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. (2023). https://doi.org/10.48550/arXiv.2310.07251 arXiv:2310.07251

[112] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1042, 18 pages. https://doi.org/10.1145/3613904.3641899

[113] Mark O. Riedl, Jonathan P. Rowe, and David K. Elson. 2008. Toward intelligent support of authoring machinima media content: story and visualization. In *Proceedings of the 2nd International Conference on INtelligent TEchnologies for Interactive EnterTAINment* (Cancun, Mexico) *(INTETAIN '08)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, Article 4, 10 pages. https://doi.org/10.4108/icst.intetain2008.2473

[114] Mark O. Riedl and R. Michael Young. 2010. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39 (Sept. 2010), 217–268. https://doi.org/10.1613/jair.2989

[115] RizenSoul. 2021. *Graph system - Dynamic lines*. Retrieved August 21, 2024 from https://assetstore.unity.com/packages/tools/utilities/graph-system-dynamic-lines-195880

[116] Yvonne Rogers and Paul Marshall. 2017. *Research in the Wild.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02220-3

[117] Jean-Jacques Rousseau. 1762/1966. *Émile ou de l'education.* Flammarion, Paris.

[118] Samantha J. Russell and Kate Cain. 2022. The animals in moral tales: Does character realism influence children's prosocial response to stories? *Journal of Experimental Child Psychology* 219 (2022), 105392. https://doi.org/10.1016/j.jecp.2022.105392

[119] Antonio Sabbatella, Andrea Ponti, Ilaria Giordani, Antonio Candelieri, and Francesco Archetti. 2024. Prompt Optimization in Large Language Models. *Mathematics* 12, 6 (2024). https://doi.org/10.3390/math12060929

[120] J. Jessica Wang Samantha J. Russell and Kate Cain. 2024. Children's Narrative Retells: The Influence of Character Realism and Storybook Theme on Central and Peripheral Detail. *Early Education and Development* 0, 0 (2024), 1–20. https://doi.org/10.1080/10409289.2024.2303908

[121] Miguel Saraiva. 2023. Oscars won by the Best Picture of the Year: An Empirical Analysis Across the History of Academy Awards (1929–2023). *Empirical Studies of the Arts* 0, 0 (2023), 02762374231212136. https://doi.org/10.1177/02762374231212136

[122] Margaret Sarlej. 2014. *A Lesson Learned: Using Emotions to Generate Stories with Morals*. Ph. D. Dissertation. Advisor(s) Ryan, Malcolm. https://doi.org/10.26190/unsworks/17061

[123] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the Moral Beliefs Encoded in LLMs. (2023). https://doi.org/10.48550/arXiv.2307.14324 arXiv:2307.14324

[124] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages. https://doi.org/10.1145/3450741.3465253

[125] Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6

(2010), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

[126] Chenxinran Shen, Yan Xu, Ray Lc, and Zhicong Lu. 2024. Seeking Soulmate via Voice: Understanding Promises and Challenges of Online Synchronized Voice-Based Mobile Dating. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 921, 14 pages. https://doi.org/10.1145/3613904.3642860

[127] Gabriel Simmons. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. (2023). https://doi.org/10.48550/arXiv.2209.12106 arXiv:2209.12106v2

[128] Abbey Singh, Ramanpreet Kaur, Peter Haltner, Matthew Peachey, Mar Gonzalez-Franco, Joseph Malloch, and Derek Reilly. 2021. Story CreatAR: a Toolkit for Spatially-Adaptive Augmented Reality Storytelling. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 713–722. https://doi.org/10.1109/VR50410.2021.00098

[129] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction* 30, 5, Article 68 (Sept. 2023), 57 pages. https://doi.org/10.1145/3511599

[130] Ho Ryun Song and Soojin Jun. 2017. Logue: Unitizing Interactive Fictions for Co-creation. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 865–876. https://doi.org/10.1145/3064663.3064761

[131] Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2023. ARN: A Comprehensive Framework and Benchmark for Analogical Reasoning on Narratives. (2023). https://doi.org/10.48550/arXiv.2310.00996 arXiv:2310.00996

[132] Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers. (2024). https://doi.org/10.48550/arXiv.2403.01061 arXiv:2403.01061

[133] Sangho Suh, Sydney Lamorea, Edith Law, and Leah Zhang-Kennedy. 2022. PrivacyToon: Concept-driven Storytelling with Creativity Support for Privacy Concepts. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (, Virtual Event, Australia,) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 41–57. https://doi.org/10.1145/3532106.3533557

[134] Suno. 2024. *"Does Suno moderate songs?"*. Retrieved November 19, 2024 from https://help.suno.com/en/articles/3198209

[135] Suno. 2024. *Suno*. Retrieved August 21, 2024 from https://suno.com/

[136] Suno. 2024. *Terms of Service*. Retrieved January 26, 2025 from https://suno.com/terms

[137] Tina M Sutton and Jeanette Altarriba. 2016. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behav. Res. Methods* 48, 2 (June 2016), 686–728. https://doi.org/10.3758/s13428-015-0598-8

[138] Cristina Sylla, Íris Susana Pires Pereira, and Gabriela Sá. 2019. Designing Manipulative Tools for Creative Multi and Cross-Cultural Storytelling. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(C&C '19)*. Association for Computing Machinery, New York, NY, USA, 396–406. https://doi.org/10.1145/3325480.3325501

[139] John R. Long STAR SYSTEMS. 2020. Aesop's Fable - Online Collection. https://aesopfables.com/

[140] Tan Tang, Sadia Rubab, Jiewen Lai, Weiwei Cui, Lingyun Yu, and Yingcai Wu. 2019. iStoryline: Effective Convergence to Hand-drawn Storylines. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 769–778. https://doi.org/10.1109/TVCG.2018.2864899

[141] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (09 2024), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

[142] Milka Trajkova, Duri Long, Manoj Deshpande, Andrea Knowlton, and Brian Magerko. 2024. Exploring Collaborative Movement Improvisation Towards the Design of LuminAI—a Co-Creative AI Dance Partner. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 890, 22 pages. https://doi.org/10.1145/3613904.3642677

[143] Carmen Tu and Steven Brown. 2020. Character mediation of plot structure: Toward an embodied model of narrative. *Frontiers of Narrative Studies* 6, 1 (2020), 77–112. https://doi.org/doi:10.1515/fns-2020-0007

[144] Maryam Vaezi and Saeed Rezaei. 2019. Development of a rubric for evaluating creative writing: a multi-phase research. *New Writing* 16, 3 (2019), 303–317. https://doi.org/10.1080/14790726.2018.1520894

[145] Tanmay H. Vakare. 2023. *Generating Morals for Fables*. Master's thesis. https://hdl.handle.net/10735.1/10037

[146] Lina Varotsi. 2019. *Conceptualisation and Exposition: A Theory of Character Construction* (1st ed.). Routledge, New York. https://doi.org/10.4324/9780429060762

[147] Caren M. Walker and Tania Lombrozo. 2017. Explaining the moral of the story. *Cognition* 167 (2017), 266–281. https://doi.org/10.1016/j.cognition.2016.11.007

[148] Qian Wan, Xin Feng, Yining Bei, Zhiqi Gao, and Zhicong Lu. 2024. Metamorpheus: Interactive, Affective, and Creative Dream Narration Through Metaphorical Visual Storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642410

[149] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 84 (April 2024), 26 pages. https://doi.org/10.1145/3637361

[150] Arran Zeyu Wang, David Borland, Tabitha C. Peck, Wenyuan Wang, and David Gotz. 2025. Causal Priors and Their Influence on Judgements of Causality in Visualized Data. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 765–775. https://doi.org/10.1109/TVCG.2024.3456381

[151] Jonah Warren. 2019. Tiny online game engines. In *2019 IEEE Games, Entertainment, Media Conference (GEM)*. 1–7. https://doi.org/10.1109/GEM.2019.8901975

[152] Joe Winston. 1995. Careful The Tale You Tell Fairy tales, drama and moral education. *Children & Society* 9, 4 (1995), 80–93. https://doi.org/10.1111/j.1099-0860.1995.tb00304.x

[153] Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5113–5125. https://doi.org/10.18653/v1/2023.emnlp-main.311

[154] Xiaotong (Tone) Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L. E, and Steven P. Dow. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 907–921. https://doi.org/10.1145/3640543.3645196

[155] Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang 'Anthony' Chen. 2023. XCreation: A Graph-based Crossmodal Generative Creativity Support Tool. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (, San Francisco, CA, USA,) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. https://doi.org/10.1145/3586183.3606826

[156] M. Yoesoef. 2020. Cyber Literature: Wattpad and Webnovel as Generation Z Reading in the Digital World. In *Proceedings of the International University Symposium on Humanities and Arts (INUSHARTS 2019)*. Atlantis Press, 128–131. https://doi.org/10.2991/assehr.k.200729.025

[157] Liane Young and James Dungan. 2012. Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience* 7, 1 (2012), 1–10. https://doi.org/10.1080/17470919.2011.569146

[158] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841—-852. https://doi.org/10.1145/3490099.3511105

[159] Sarfaroz Yunusov, Hamza Sidat, and Ali Emami. 2024. MirrorStories: Reflecting Diversity through Personalized Narrative Generation with Large Language Models. (2024). https://doi.org/10.48550/arXiv.2409.13935 arXiv:2409.13935

[160] Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 620, 25 pages. https://doi.org/10.1145/3613904.3642877

[161] Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making* 18 (2023), e41. https://doi.org/10.1017/jdm.2023.37

[162] Yangsheng (Danson) Zheng and Nicola Stewart. 2024. Improving EFL students' cultural awareness: Reframing moral dilemmatic stories with ChatGPT. *Computers and Education: Artificial Intelligence* 6 (2024), 100223. https://doi.org/10.1016/j.caeai.2024.100223

[163] David Zhou and Sarah Sterman. 2024. Ai.llude: Investigating Rewriting AI-Generated Text to Support Creative Expression. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 241–254. https://doi.org/10.1145/3635636.3656187

[164] Fabio Zünd, Steven Poulakos, Mubbasir Kapadia, and Robert W. Sumner. 2017. Story Version Control and Graphical Visualization for Collaborative Story Authoring. In *Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017)* (London, United Kingdom) *(CVMP '17)*. Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages. https://doi.org/10.1145/3150165.3150175

A LLM Selection Details

A.1 Demographics

Participant demographic information can be found in Table 6.

A.2 Prompts

To generate morals of the story (i.e., for 1) in Section 4), we used the prompt: "What is the moral of the story based on the OUTLINE below? Answer in a single sentence." For outlines (i.e., 2) in Section 4), based on formative participant feedback and story coherence, we used the prompt "Change the plot (not just the wording) of the OUTLINE below to better reflect the intended MORAL while maintaining the initial setting and the writer's style. Keep the articles ('a,' 'an,' 'the') used in the original OUTLINE the same. The number of sentences needs to be between 5-10. Each sentence is a story event. Do not state the MORAL explicitly if it hasn't been stated in the original OUTLINE."

A.3 Results

In total, for each of 1) and 2) (Section 4), the evaluation has 9 questions (order randomized) and 198 comparisons for each condition pair. When comparing with human outputs, evaluators chose LLMs/"Similar" more often (for 1) GPT/human: 103 GPT and 40 "Similar"; for 1) Claude/human: 99 Claude and 40 "Similar"; for 2) GPT/human: 102 GPT and 34 "Similar"; for 2) Claude/human: 102 Claude and 26 "Similar"). For GPT/Claude, evaluators chose GPT for 70 comparisons for 1) and 90 for 2) and chose Claude for 71 for 1) and 69 for 2).

B External Evaluation Details

B.1 Demographics

Participant demographic information can be found in Table 7.

B.2 Response Pairs Evaluated

For the first part of the questionnaire, each evaluator evaluated all response pairs described in Table 8.

B.3 Preferences for Information Presentation in an Event Node Graph

Questions about preferences for information presentation in external evaluators' questionnaire are a series of comparisons starting with the following: "Assume you have to review an outline similar to the ones you have just read. This outline is presented in the form of an event node graph, a graph where each story event (each paragraph in the outlines you have read) is within a node and the nodes are connected to each other with an arrow indicating the logical progression of the story. The graph and the outline are by someone else, so you did not know how the graph would look before you get it. You also didn't know what happens in the story. Which of the following event node graphs more clearly presents the story to you?"

The comparisons and choices are as follows. For "all the same node color and shape" versus "nodes with the same color but different shapes to represent different types of information", 15 chose *the second* and 4 *"It depends. If the nodes are colored based on how I would color them, they can help me understand the story more quickly. If the nodes are colored differently, even if there are instructions, it takes more effort to develop understanding. So, it can be distracting."* For "all the same node color and shape" versus "nodes with the same color but different shapes to represent different types of information", 1 chose *the first*, 10 *the second*, 7 *"It depends. If the nodes are shaped based on how I would shape them, they can help me understand the story more quickly. If the nodes are shaped differently, even if there are instructions, it takes more effort to develop understanding. So, it can be distracting."* and 1 *"It doesn't matter. When I look at a node graph, shapes as visual cues are irrelevant to me."* For "nodes with the same shape but different colors to represent different types of information" versus "nodes with the same color but different shapes to represent the same types of information", 9 chose *the first*, 3 *the second*, and 5 *"A combination (using both colors and shapes) would be the clearest."* For "same size for both the node containing the short description of the event and for the node containing details about the event (ASSUMING YOU CAN ZOOM IN AND ZOOM OUT WITHOUT THE TEXT BECOMING BLURRY)" versus "smaller size for the node containing details about the event (ASSUMING YOU CAN ZOOM IN AND ZOOM OUT WITHOUT THE TEXT BECOMING BLURRY)", 6 chose *the first*, and 13 *the second*.

Example images are provided for each type of graph with the warning "Note: there are many ways to construct such a graph. This is only one possibility to give you an idea. Please mainly rely on the text description of the characteristics being compared." An example pair of image examples can be found in Figure 12.

ID	Age	Gender	Locations	Professional Experience	Education
S1	28	Female	CN	P	Formal (classes)
S2	30	Female	CN, CN (HK)	P	Formal (workshops)
S3	19	Male	CN, CN (HK), Japan	P	Formal (classes)
S4	26	Male	CN, CN (HK), Germany, Japan, Malaysia, Papua New Guinea, US	P	Formal (classes)
S5	24	Male	CN	None	Informal
S6	19	Male	CN, CN (HK), Cuba, France, UK, US	P	Informal
S7	24	Female	Austria, Canada, CN, CN (HK), France, Germany, India, Italy, Morocco, Switzerland, US	None	Formal (classes)
S8	21	Male	Brazil, CN, CN (HK), CN (M), Egypt, India, Indonesia, Japan, Myanmar, South Korea, Thailand	None	Formal (workshops)
S9	19	Female	CN	P	Informal
S10	19	Male	CN, Japan	None	Formal (classes)
S11	22	Female	CN	None	Formal (workshops)
S12	25	Male	Albania, Canada, CN, CN (HK), Germany, Israel, Japan, Mexico, Singapore, US	P	Formal (classes)
S13	22	Non-binary / third gender	Afghanistan, Argentina, Bosnia and Herzegovina, Chile, CN, CN (HK), Czech Republic, France, Germany, Iceland, Iran, Italy, Japan, Mexico, Netherlands, Portugal, Romania, Russian Federation, South Korea, Spain, UK, US, Viet Nam	FP	Formal (classes)
S14	19	Male	CN	None	Informal
S15	20	Male	India	P	Informal
S16	24	Male	CN	P	Informal
S17	30	Female	CN, US	None	Formal (classes)
S18	57	Female	Canada, CN, CN (HK), France, Japan, South Korea, Spain, Thailand, US	P	Formal (classes)
S19	24	Male	CN	P	Formal (degree)
S20	29	Male	Australia, Austria, CN, CN (HK), Japan, US	FP	Formal (degree)
S21	26	Male	CN, Japan, Philippines, US	P	Formal (classes)
S22	29	Male	CN	FP	Formal (degree)

Table 6: Demographic information of evaluators for the LLM selection. For "Gender", 7 chose "Female", 14 chose "Male", and 1 chose "Non-binary/third gender". Evaluators are aged from 19 to 57 years old, with an average of 25.3. We use the same abbreviations for "Locations", "Professional Experience", and "Education" as Table 3. "CN (M)" means *Macau (S.A.R. China).*

ID	Age	Gender	Locations	Professional Experience	Education
E1	24	Male	CN, CN (HK), CN (M), Indonesia, Malaysia, Russian Federation	None	Informal
E2	18	Female	CN	None	Informal
E3	22	Female	CN	P	Informal
E4	20	Female	CN, France, Japan	None	Informal
E5	25	Male	CN, CN (HK)	P	Formal (degree)
E6	31	Male	CN	FP	Informal
E7	24	Female	CN, Japan, US	P	Formal (classes)
E8	22	Male	CN	None	Formal (classes)
E9	25	Male	CN	None	Informal
E10	24	Female	CN	P	Formal (degree)
E11	23	Male	CN, CN (HK), Germany	P	Formal (classes)
E12	19	Male	CN	None	Formal (classes)
E13	25	Male	Albania, Brazil, Canada, CN, CN (HK), Germany, India, Israel, Japan, Malaysia, Mexico, Russian Federation, Singapore, South Korea, UK, US	P	Formal (classes)
E14	19	Female	CN (HK)	None	Formal (classes)
E15	24	Male	CN, CN (HK), US	P	Informal
E16	22	Male	CN	None	Formal (classes)
E17	23	Male	CN (HK)	F	Formal (classes)
E18	21	Male	CN, CN (HK), CN (M), France, Indonesia	None	Informal
E19	18	Male	Belarus, Canada, CN, CN (HK), Cuba, Finland, Iceland, Israel, Japan, Mongolia, Singapore, Ukraine, US	None	Informal

Table 7: Demographic information of evaluators for the external evaluation of user study task responses. For "Gender", 6 chose "Female" and 13 chose "Male". Evaluators are aged from 18 to 31 years old, with an average of 22.6. We use the same abbreviations for "Locations", "Professional Experience", and "Education" as Table 3. "CN (M)" means *Macau (S.A.R. China).*

Moral	**Audience**
Love is not only about happiness but also deep understanding of each other.	those who have had experience with intimate relationships
In family relationships, there should be mutual understanding because parents are also first-time parents.	kids and parents
Keep practicing or you may miss opportunities because you are not well prepared.	kids
Greed and disunity can lead to one's downfall.	teenagers
Any act may be justified by the degree of positive change it brings about.	teenagers (those who are into social media)
Death is not the end of a person's life story but rather another starting point.	young professionals struggling with existential crisis
Humans should not be too arrogant; all life is equal. In the eyes of higher beings, humans are nothing more than that.	young kids
Even in chaotic and unpredictable situations, mutual understanding, cooperation, and empathy can lead to unexpected friendships and solutions.	a 60-year-old woman in hospital
Integrity and hard work will always shine brighter than any shortcut.	children in primary school
True friendship transcends backgrounds and circumstances.	kids in kindergarten
It's never too late to pursue new interests and share your passions, which can lead to personal fulfillment and community building.	old adults who have retired
True fulfillment and happiness often come from following one's passions and making a positive impact on society, rather than merely accumulating wealth.	young graduate students
Being always immersed in the past is meaningless. We need to focus on what we have and what we can do in the present.	those who focus on the past, who focus on what they lost and what they suffer from
Finding a balance in an indulgence-abundant and stressful world is important.	people severely craving indulgence and people who live without any indulgence
Human activities destroy the nature, and the grassroots are trying to fight against the monopoly.	a movie director who writes Sci-Fi movies
Embrace your passions and overcome your fears to find true fulfillment and inspire others.	a movie director for romance stories
Nobody can always get things right; obstacles are meant to be learning experiences. "What doesn't kill you makes you stronger."	people interested in magical creatures, adventure
Even individuals can have an impact (grassroots power).	young people
Identity is loose and changing through life. Acceptance of other forms, shapes, and ways of being and self-acceptance of what we know ourselves to be are important.	any age
The moral of the story revolves around the immense power and responsibility of collective human thought and the consequences of using such power unethically.	culturally diverse young adults who are often more open to exploring self-improvement and spiritual practices
Helping others is good.	adults
All for one's own benefits.	adults
We should pursue a long-term vision instead of focusing on certain quantifiable achievements.	high school students who are facing university entrance examination pressures
True success comes from staying true to your values, fostering effective communication, and focusing on quality and craftsmanship, rather than chasing fleeting trends.	young elephants as grandsons and daughters of the elephant's chocolate company, who are facing the age of AI
Advancement of technology may lead to lack of meaning in people's lives.	general public
Live in the present and not dwell on past regrets or try to manipulate the future.	young adults (18-30): people at a stage where they're making important life decisions and shaping their futures
This story tells people to be good at observing the details of life, to understand the people and things around them, and not to focus only on themselves.	kids in elementary school who love to read. We want to help them develop moral values through reading.
The moral of the story is that collaboration and embracing different perspectives can lead to personal growth and success in artistic endeavors.	recent college graduates who feel lost in life

Table 8: The moral of the story and the target audience for each user study response pair.

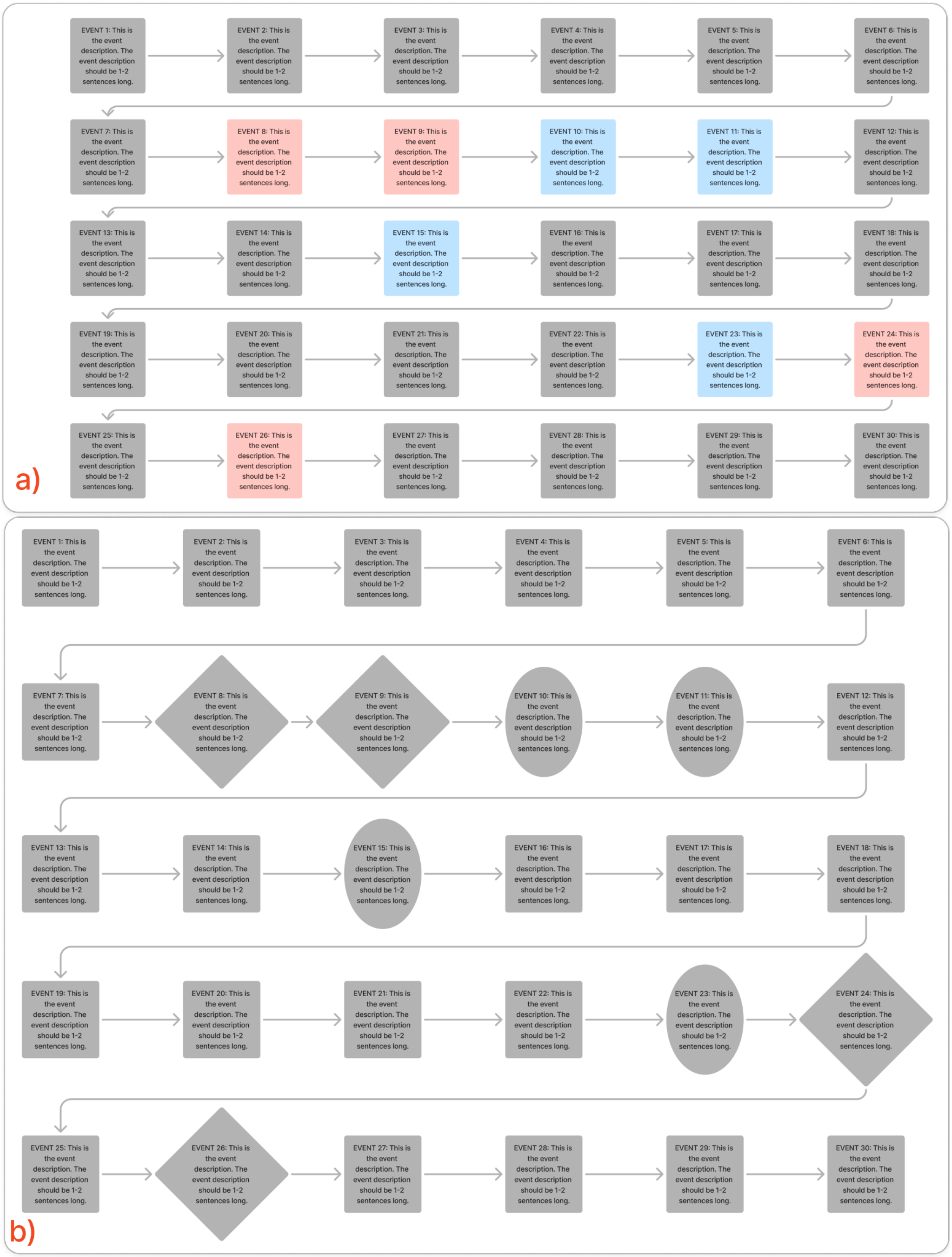

Figure 12: Example images for the question asking to compare graphs with a) "nodes with the same shape but different colors to represent different types of information" versus with b) "nodes with the same color but different shapes to represent the same types of information" in external evaluators' questionnaire (Section B.3).