NarrativeLoom: Enhancing Creative Storytelling through Multi-Persona Collaborative Improvisation

Yuxi Ma[⋆123]
Institute for Artificial Intelligence, Peking University
Beijing, China
yxma@stu.pku.edu.cn

Yongqian Peng[⋆124]
Institute for Artificial Intelligence, Peking University.
Beijing, China
yqpeng@stu.pku.edu.cn

Fengyuan Yang[124]
Institute for Artificial Intelligence, Peking University
Beijing, China
fyyang@stu.pku.edu.cn

Siyu Zha
The Future Laboratory, Tsinghua University
Beijing, China
zhasiyu22@mails.tsinghua.edu.cn

Chi Zhang[2]
School of Intelligence Science and Technology, Peking University
Beijing, China
chizhang.cz@pku.edu.cn

Zixia Jia[2]
Beijing Institute for General Artificial Intelligence (BIGAI)
Beijing, China
jiazixia@bigai.ai

Zilong Zheng[✉2]
Beijing Institute for General Artificial Intelligence (BIGAI)
Beijing, China
zlzheng@bigai.ai

Yixin Zhu[✉235]
School of Psychological and Cognitive Sciences, Peking University
Beijing, China
yixin.zhu@pku.edu.cn

Project Website: https://ppyyqq.github.io/improviser/

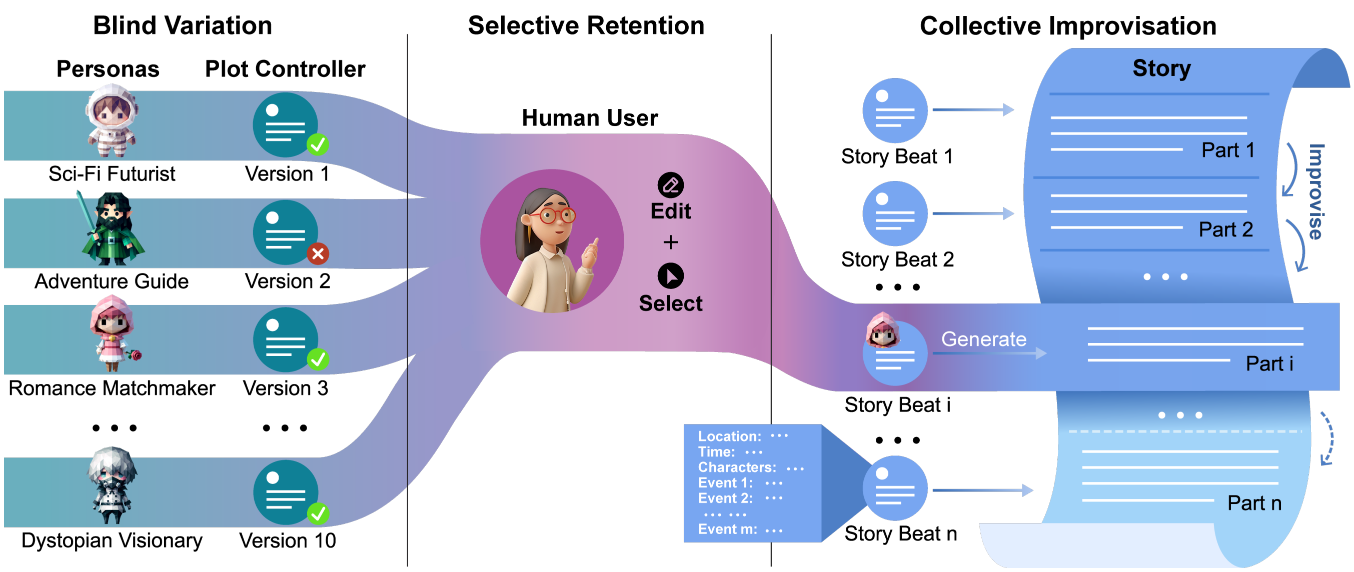


Figure 1: The architecture of NarrativeLoom, the multi-persona co-creation system inspired by the Blind Variation and Selective Retention (BVSR) theory. Our computational implementation of BVSR operates through three interconnected phases: (i) *Blind Variation*: Ten specialized storyteller personas independently generate diverse narrative possibilities for each story beat, with a plot controller ensuring baseline coherence through consistency verification; (ii) *Selective Retention*: The human user evaluates, selects, and optionally edits the most promising beat from the generated alternatives, exercising creative direction; and (iii) *Collective Improvisation*: Selected story beats—containing structured narrative elements (location, time, characters, events)—are sequentially transformed into cohesive narrative segments that progressively build the complete story through collaborative human-AI iteration.

Abstract

Large Language Models show promise for AI-assisted storytelling, yet current tools often generate predictable, unoriginal narratives. To address this limitation, we present NarrativeLoom, a multi-persona co-creative system grounded in Campbell's Blind Variation and Selective Retention (BVSR) theory. NarrativeLoom deploys specialized Artificial Intelligence (AI) personas to generate diverse narrative options (blind variation), while users act as creative directors to select and refine them (selective retention). We designed a controlled study with 50 participants and found that stories co-authored with NarrativeLoom were not only perceived by users as more novel and diverse but were also objectively rated by experts as significantly better across all Torrance Test creativity dimensions: fluency, flexibility, originality, and elaboration. Stories are significantly longer with richer settings and more dialogue. Writing expertise emerged as a moderator: novices benefited more from structured scaffolding. This demonstrates the value of theory-informed co-creative systems and the importance of adapting them to varying user expertise.

CCS Concepts

• **Human-centered computing → Systems and tools for interaction design**; **Collaborative interaction**; *Collaborative and social computing systems and tools*; **Human computer interaction (HCI)**.

Keywords

Human-computer interaction, Large language models, Storytelling, Creative support tool

ACM Reference Format:

Yuxi Ma, Yongqian Peng, Fengyuan Yang, Siyu Zha, Chi Zhang, Zixia Jia, Zilong Zheng, and Yixin Zhu. 2026. NarrativeLoom: Enhancing Creative Storytelling through Multi-Persona Collaborative Improvisation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3790416

1 Introduction

Storytelling is a fundamental pillar of human culture and a primary medium for sharing knowledge, values, and experiences [39, 58]. Effective narratives, from Aristotle's *Poetics* [4] to modern screenplays, depend on a balance between coherence and surprise. They require "peripeteia"—unexpected turns that engage audiences not through randomness, but by revealing a deeper, latent logic [26]. This combination of novelty and appropriateness characterizes human creativity and remains a central focus of computational creativity research [28, 50].

Large Language Models (LLMs) such as GPT-4 [1] and Gemini [60] have changed the landscape of computational narrative generation. These models provide higher fluency and contextual understanding than earlier rule-based systems [2, 63]. However, the Human-Computer Interaction (HCI) community has noted that current AI storytelling systems often remain conservative [10, 16, 43, 68]. Because their architectures are optimized for next-token prediction, they excel at producing statistically probable continuations but often suppress the surprising deviations essential for creative depth [10, 11, 26, 28]. Industry professionals have observed that this limitation [3, 15, 16, 19, 43] often results in predictable, cliché narratives [11, 35, 50, 67], leading to homogenized ideas.

Our formative study investigating how writers experience these limitations revealed a tension between the desire for creative diversity and the need for narrative control. Current tools often fail to resolve this; they either suggest predictable paths or generate content disconnected from the writer's vision. This highlights a key HCI challenge: enabling LLM-based systems to generate meaningful surprises while maintaining user agency.

To address these needs, we applied Campbell's theory of Blind Variation and Selective Retention (BVSR) [8]. BVSR describes a two-phase process: generating unconstrained ideas ("blind variation") followed by the deliberate curation of promising ones ("selective retention"). This framework serves as a theoretical blueprint for a system that offers a broader set of possibilities while preserving the user's creative authority. Based on this foundation, we developed NarrativeLoom, a system that operationalizes BVSR through multi-persona collaborative improvisation [52, 54]. To implement "blind variation," the system employs an ensemble of specialized AI personas, each providing a unique genre-aware narrative lens to generate diverse story beats. To facilitate "selective retention," users act as creative directors who select, edit, and integrate these beats. Our approach differs from conventional tools by: (i) separating generative and curatorial processes; (ii) achieving variation through specialized personas rather than parametric sampling; and (iii) scaffolding exploration while maintaining user authority.

Our findings demonstrate that NarrativeLoom's BVSR-based, multi-persona approach significantly enhances creative outcomes compared to the single-voice chatbot, producing longer, richer stories with more settings and higher dialogue ratios that emphasize "showing over telling [33]." Users perceived NarrativeLoom as providing more diverse narrative possibilities while maintaining high usability, with strategic persona engagement revealing asymmetric transition patterns where certain personas serve as "initiators" and others as "developers." Professional writing experts rated stories generated by NarrativeLoom significantly higher across all creativity dimensions—fluency, flexibility, originality, and elaboration—praising its ability to create unexpected narrative turns and psychologically complex characters. Writing experience emerged as a key

★Both authors contributed equally to this research.
[1]Also with the School of Psychological and Cognitive Sciences, Peking University.
[2]Also with the State Key Lab of General AI.
[3]Also with the Beijing Key Laboratory of Behavior and Mental Health, Peking University.
[4]Also with Yuanpei College, Peking University.
✉Corresponding authors
[5]Also with Institute for Artificial Intelligence, Peking University

CHI '26, April 13–17, 2026, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790416

moderating factor, with novice writers benefiting more from NarrativeLoom's structured scaffolding, suggesting particularly strong support for writers developing narrative intuitions.

This research offers three primary contributions to the fields of HCI and computational creativity: (i) a **theoretical framework** that operationalizes Campbell's Blind Variation and Selective Retention (BVSR) theory within human-AI co-authorship, providing a principled mechanism to balance AI-driven diversity with human executive control; (ii) a **system design** utilizing a multi-persona collective improvisation architecture, which structures creative variation through the functional heterogeneity of specialized personas rather than stochastic parametric sampling; and (iii) **empirical insights** revealing that both writing expertise and creative stages significantly moderate system effectiveness, necessitating adaptive interfaces that accommodate varying skill levels and shifting creative contexts.

2 Related Work

2.1 Human-AI Co-creation in Storytelling

Human-AI co-creation in storytelling has shifted from basic creativity support to collaborative partnerships that redistribute agency between writers and computational systems. Recent HCI research examines the social and cognitive dynamics of these interactions. For instance, Gero et al. [27] shows how writers navigate the boundary between treating AI as a tool *vs.* a collaborator, while Dhillon et al. [19] finds that different scaffolding levels alter both the creative process and narrative quality. Although LLMs provide high fluency, professional writers report concerns regarding creative homogenization and the difficulty of maintaining an authentic voice within systems like Wordcraft [35, 68].

To address these limitations, the HCI community has developed interaction paradigms that offer control beyond text generation. Chung et al. [15] introduced TaleBrush, which uses line-level sketching to shape story arcs, demonstrating that visual modalities can effectively direct creative outcomes. Addressing the need for structural control, Lu et al. [41] proposed WhatELSE, a system that allows authors to manage narrative abstraction through planning to ensure causal soundness in open-ended plots. Similarly, Mirowski et al. [43] argues for tools that serve as creative partners rather than generators in screenplay writing, while Chakrabarty et al. [11] investigated iterative editing interfaces to align human intent with model output.

Despite these innovations, maintaining structural coherence in long-form narratives remains difficult. While hierarchical frameworks use top-down planning to address this [66], Mirowski et al. [43] argue that rigid structures conflict with the emergent nature of writing, often producing predictable patterns. Chakrabarty et al. [10] characterize this as a "false promise of creativity," where models optimized for next-token prediction fall into a "probability trap" [26]. This results in technically proficient but conservative content that lacks the deviations necessary for compelling storytelling. These tensions between coherence and diversity, and between agency and assistance, indicate a need for approaches that support narrative variety while preserving authorial control.

2.2 Improvisational Storytelling

Improvisational storytelling involves the construction of spontaneous narratives without predetermined elements [52, 54]. This process requires continuous adaptation and generates uncertainty, distinguishing it from hierarchical approaches [29, 59]. This framework is aligned with the distributed creativity theory [53, 54], which posits that creativity emerges from interactional dynamics rather than isolated acts. The HCI community has applied these principles to collaborative systems; for example, Kim et al. [38] developed the Ensemble system, which asymmetrically distributes creative responsibility between leaders and crowds.

In Ensemble [38], lead authors maintain a high-level vision through "scene prompts" while contributors generate content within specified boundaries. This suggests that strategic constraints can focus attention on specific narrative elements to support creative freedom. Our system extends this principle by using story beats as navigational constraints and replacing human crowds with specialized AI personas. This configuration shifts the leader-contributor dynamic into a human-AI partnership mediated by specific creative voices, allowing for improvisational emergence within a structured framework.

Recent research has examined AI capabilities in distributed creative processes. Wang et al. [64] investigated "cognitive synergy" through Solo Performance Prompting (SPP), where a single LLM adopts multiple collaborative personas. Their results indicated that using specialized personas reduced hallucinations and maintained reasoning capabilities during complex tasks. While SPP was evaluated on task-solving rather than narrative generation, it provides a computational basis for applying distributed creativity to collaborative improvisational storytelling.

3 Formative Study

To understand the challenges in human-AI collaborative storytelling, we conducted a formative study involving in-depth interviews with writers of varying expertise. Our objective was to identify specific areas where writers require support, thereby informing the design of an AI-based system for collaborative narrative creation.

3.1 Methods

3.1.1 Participants. Writers are the primary practitioners of narrative storytelling across different media. We recruited five participants through professional networks to ensure a range of expertise, from emerging writers (2 years) to experienced professionals (15+ years; see Tab. 1). All participants had used AI writing tools previously, primarily for brainstorming, research, and story development.

3.1.2 Procedure. We conducted semi-structured remote interviews via video conferencing, each lasting approximately one hour. The protocol examined participants' experiences across story development phases: ideation, planning, drafting, and revision. We investigated current AI tool usage, specific strategies, challenges, and perspectives on how AI could support creative storytelling. Participants discussed collaborative writing experiences and their expectations for maintaining agency in human-AI partnerships.

Table 1: Participant demographics and professional background. This table details the gender, age, occupation, education, and years of writing experience for the five study participants.

ID	Gender	Age	Occupation	Education	Writing Experience (Years)
W1	Male	34	Screenwriter	Master of Fine Arts in Film Production	8
W2	Male	34	Writer	Master of Arts in Creative Writing	15+
W3	Female	38	Film Producer/Screenwriter	Bachelor of Arts in Literature	10+
W4	Female	27	Freelance Writer	Master of Arts in Anthropology	2
W5	Male	32	Content Creator/Educator	Bachelor of Medicine	7

3.2 Key Findings

Analysis revealed four areas where writers seek support in collaborative storytelling:

Managing Narrative Structure through Segmented Units. Participants consistently struggled with managing coherence and creative momentum across extended narratives, particularly when balancing spontaneous creativity with structural organization. W1 described the creative process as involving *"spiral progression,"* where writers *"repeatedly break the simple linear, top-down structure,"*. W4 emphasized that compelling writing contains *"randomness"* that *"cannot be explained by high-probability experiences,"* suggesting that effective structural support must accommodate unexpected creative developments. Writers naturally adopted segmented approaches to manage this complexity. W2 utilized *"story beats"* as discrete structural units for maintaining narrative progression without constraining exploration. W1 noted that while *"short stories don't require such strong structural demands,"* longer works benefit from *"breaking it down into familiar story structures like three-act or eight-sequence formats"* that provide navigational waypoints without dictating creative content.

Seeking Diverse Perspectives Beyond Single-Voice Generation. All participants reported creative limitations with single-voice AI systems, perceiving them as producing repetitive content lacking genuinely novel narrative elements. W2 explained that existing AI systems *"just continue what you're doing"* rather than bringing *"new beats or new elements"* to the story. Writers valued exposure to multiple perspectives during the creative process. W5 advocated for *"multiple viewpoints"* from different disciplinary backgrounds, while W3 emphasized that individual creative capacity is inherently *"limited"* and benefits from diverse input. W1 described AI's potential strength in *"finding a partner"* that could provide *"different possibilities"* and perform creative *"combinations,"* suggesting effective AI storytelling systems should provide diverse creative alternatives.

Retaining Creative Ownership during AI Assistance. While appreciating AI assistance, participants strongly emphasized maintaining creative ownership and preserving improvisational creativity. W3 articulated this need clearly: *"I am the one making judgments, I am the one making the final decisions... so this is my story."* However, participants criticized inefficient interaction patterns with current AI tools. W2 described the frustration with *"continually prompting it for like half an hour"* without productive outcomes, highlighting the need for more effective collaborative workflows that preserve creative agency while providing needed support.

Coordinating Consistency across Multi-Voice Contributions. A challenge emerged when managing coherence across multiple creative contributors. W2 highlighted collaboration difficulties, noting how different writing styles can result in *"clashing with each other."* When multiple voices contribute to a single narrative, maintaining consistent character development, plot logic, and thematic coherence becomes increasingly complex and requires careful coordination beyond individual writing processes. W3 observed AI's limitations in maintaining consistency across extended collaborative works, describing current systems as having limited capacity for coherent long-form generation. This challenge encompasses more than basic fact-checking—it involves coordinating deeper narrative elements, including character development arcs, thematic consistency, and tone maintenance throughout extended collaborative creation processes.

3.3 Design Goals

Based on these findings, we identified four design goals for human-AI collaborative storytelling:

DG1: Structuring Creative Development through Narrative Units. The system should decompose story creation into discrete, well-defined narrative units (e.g., story beats). These serve as creative waypoints, allowing writers to balance improvisation with structural organization in extended narratives.

DG2: Expanding Creative Exploration through Diverse Narrative Voices. To overcome the repetitive nature of single-model generation, the system should employ multiple specialized creative voices. These voices should provide unexpected directions and diverse perspectives to enhance exploration.

DG3: Empowering User Agency through Selective Control. The system should position users as creative directors who evaluate, select, and refine AI-generated elements. Mechanisms should enhance human decision-making rather than automate the creative process entirely.

DG4: Supporting Narrative Coherence across Collaborative Inputs. The system should monitor and support narrative coherence across diverse contributions. It must help writers identify and resolve inconsistencies in character arcs and thematic elements to ensure quality throughout the collaborative process.

4 System Design

4.1 Theoretical Foundation

Based on our formative study, we identified the need for a framework that balances creative diversity with human agency and narrative coherence. We adopted Campbell's Blind Variation and Selective Retention (BVSR) model [8], which applies evolutionary principles to creativity. BVSR posits that creative processes require two distinct phases: the generation of diverse alternatives (blind variation) and the systematic evaluation and retention of promising options (selective retention).

Blind Variation: This phase involves generating alternatives independently of existing patterns or statistical likelihoods. This prevents the system from converging on predictable outputs, which is a known limitation of next-token prediction in LLMs. By implementing variation through specialized personas, the system generates diversity at structural and causal levels rather than only varying surface-level linguistics. This approach directly supports *Design Goal 2*.

Selective Retention: In this phase, promising variations are evaluated and retained. Because this requires contextual understanding and domain expertise, the system assigns this role to the user. This maintains human agency by positioning the user as the primary decision-maker, supporting *Design Goal 3*. Within this framework, the AI provides generative support while the human user guides the narrative trajectory.

Iterative Cycles: BVSR suggests that creativity emerges through repeated variation-selection cycles. Each selection provides the narrative context for subsequent variations. This informs our beat-based architecture *(Design Goal 1)*, where the system generates multiple variations for each story beat, and human selection guides narrative progression. This structure supports *Design Goal 4* by enabling coherence through the interplay between algorithmic variation and human curation, rather than relying solely on predefined narrative constraints.

4.2 Design Principles

This section details the design decisions used to implement BVSR theory within a co-creative storytelling system.

4.2.1 User Workflow Design. NarrativeLoom uses a three-phase workflow (see Fig. 2) based on **story beats**—discrete units of narrative progression containing settings, characters, and events. In screenwriting, beats function as the fundamental components of narrative arcs [44], making them an effective unit for variation-selection dynamics.

Discovery and Ideation: Users provide initial narrative inputs ("sparkles") and define story parameters. This establishes the initial state while preserving a broad exploration space for subsequent variation.

Collaborative Story Creation: This phase implements variation through multi-persona generation. Ten specialized personas generate beat alternatives simultaneously. The interface displays the structural details and rationale for each proposal. Users then perform selective retention by evaluating these alternatives against their narrative goals. Users can modify selected beats before expanding them into 800–1000 word prose.

Iteration: Users repeat the generation and selection process to build the narrative. The system integrates previous context while maintaining generative diversity. A RAG-based consistency system encodes story history into semantic embeddings. When new beats are generated, this mechanism identifies logical inconsistencies and adjusts rankings to prioritize coherent options while preserving alternatives.

4.2.2 Story Beat Architecture [DG1].

Design Rationale. To maintain manageability, storytelling is structured into discrete *story beats*. Our design prioritizes semantic transparency by decomposing each narrative segment into setting, characters, and key events. This modularity allows for precise modifications—such as changing a specific character action—without requiring the regeneration of the entire scene. This granularity provides clear causal anchors for the model and ensures contextual coherence.

4.2.3 Multi-Persona Generation for Creative Diversity [DG2].

Design Rationale. To increase output diversity, the system uses genre-based rather than style-based personas. Style-oriented approaches primarily affect lexical choice, whereas genre-based personas influence narrative logic and causal structures. Genre dictates the types of events and the underlying logic of the narrative. For example, a *Mystery* persona integrates structural elements such as information asymmetry and the strategic placement of clues.

Persona Design. We selected ten personas based on three criteria:

- **Genre Coverage**: Representation across major narrative categories grounded in established literary frameworks [21].
- **Narrative Differentiation**: Personas span different approaches, including plot-driven (*e.g.*, *Adventure Guide*), character-centric (*Romance Matchmaker*), and world-building roles (*Fantasy World Builder*).
- **Complementary Functions**: Personas emphasize different narrative elements, such as atmosphere (*Horror Atmosphere Creator*) or social commentary (*Dystopian Visionary*).

This diversity enables NarrativeLoom to explore different regions of the narrative space simultaneously, rather than clustering around a single narrative trajectory with minor variations. The complete persona specifications are detailed in Tab. 2.

4.2.4 Selective Retention Interface for Creative Agency [DG3].

Design Rationale. The interface positions users as the final evaluators through a multi-layered design. This supports different levels of engagement based on the user's specific creative goals.

Interaction Design. NarrativeLoom preserves agency through three patterns:

- **Direct Selection:** This design supports a fluid drafting flow by allowing writers to adopt preferred beats with a single click. This mode is designed for moments of high creative momentum, where the priority is rapid progression rather than granular deliberation.
- **Comparative Evaluation:** This design encourages reflective decision-making through side-by-side comparison. To minimize cognitive load, the interface presents two alternatives simultaneously while providing access to full candidates through progressive disclosure.
- **Multi-stage Editing:** This design treats AI-generated suggestions as provocative starting points rather than finalized text. This design allows writers to refine both individual story beats and expanded narratives, ensuring that the final output remains grounded in the author's unique voice and intent.

4.2.5 Consistency Management Through Soft Constraints [DG4].

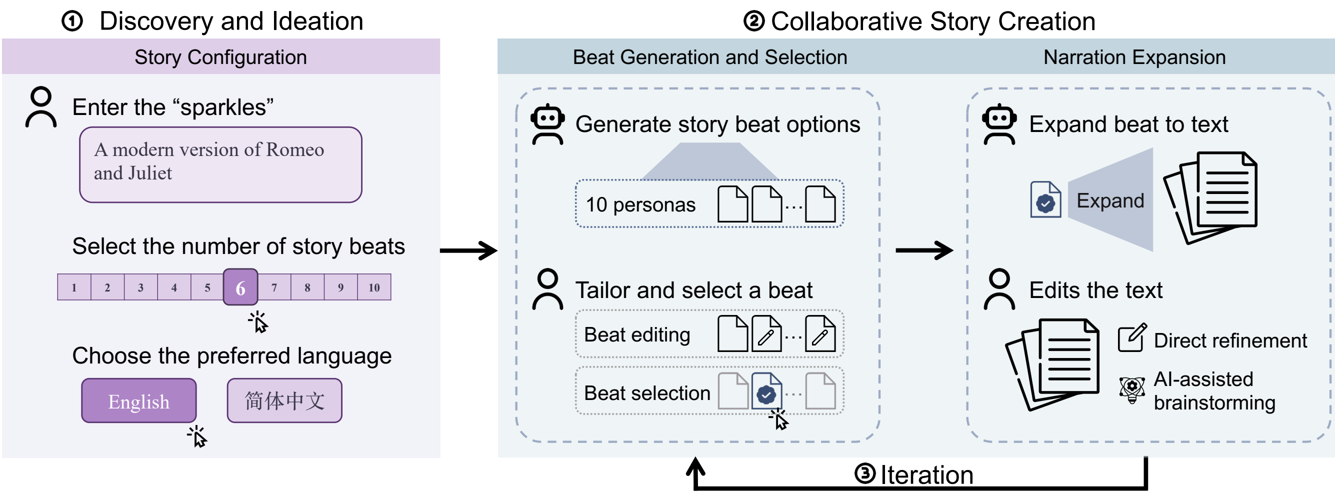


Figure 2: The user workflow of NarrativeLoom. The process consists of three integrated phases: (i) Discovery and Ideation, where users initialize the narrative by entering "sparkles" and selecting parameters such as language and story length; (ii) Collaborative Story Creation, where the system generates diverse beat options using 10 distinct personas, allowing users to tailor, select, and expand beats into full narrative, and edit the text before iterating to the next story beat; and (iii) Iteration, where users build their story progressively by repeating the beat selection and narrative expansion process.

Table 2: Storyteller personas in NarrativeLoom. Ten specialized storytelling personas form the generative ensemble of NarrativeLoom, each designed with distinct genre expertise and narrative capabilities.

Persona	Narrative Specialization
Fantasy World Builder	Specialize in crafting rich and imaginative fantasy worlds, complete with intricate magic systems, mythical creatures, and diverse cultures.
Sci-Fi Futurist	Focus on creating believable and innovative science fiction settings, incorporating advanced technology, space travel, and futuristic societies.
Mystery Solver	Assist in developing complex and intriguing mysteries, helping to plant clues, red herrings, and plot twists that keep readers engaged until the narrative resolution.
Romance Matchmaker	Skilled at creating compelling romantic storylines, ensuring that character chemistry feels authentic and that relationships develop naturally over narrative progression.
Historical Researcher	Excel at incorporating accurate historical details and context into narratives, bringing historical fiction to life and immersing readers in specific temporal settings.
Horror Atmosphere Creator	Help to build tension and suspense in horror narratives, using descriptive language and pacing to create unsettling atmospheric elements that enhance reader engagement.
Adventure Guide	Specialize in crafting thrilling adventure stories, designing exciting action sequences, perilous obstacles, and high-stakes challenges for character development.
Comedy Humorist	Focus on incorporating humor and wit into narratives, using wordplay, situational comedy, and character interactions to enhance narrative enjoyment.
Dystopian Visionary	Adept at constructing dystopian settings and exploring societal and political implications, helping to create thought-provoking and cautionary narrative frameworks.
Magical Realism Conjuror	Assist in blending fantastical elements with everyday reality, creating narratives that are simultaneously grounded and whimsical, featuring extraordinary occurrences within otherwise ordinary contexts.

Design Rationale. Managing coherence requires balancing narrative stability with creative variation. Rather than using hard constraints that automatically filter out inconsistent options, we implemented a ranking system. This prioritizes logically consistent paths while allowing users to choose divergent options if they serve a specific creative purpose.

4.3 Technical Implementation

4.3.1 Beat Generation Pipeline.

Technical Architecture. The beat generation employs a three-layer prompt architecture (see Fig. 3). The **meta-prompt layer** establishes the collaborative storytelling framework for the multi-persona system. The **context integration layer** combines compressed story history with the current beat state to maintain narrative continuity. The **generation constraint layer** specifies structural requirements and coherence criteria for outputs.

Data Schema and Structured Output. Each story beat is generated as a structured JSON object with three key fields: `setting` for spatio-temporal context, `characters` for active participants, and

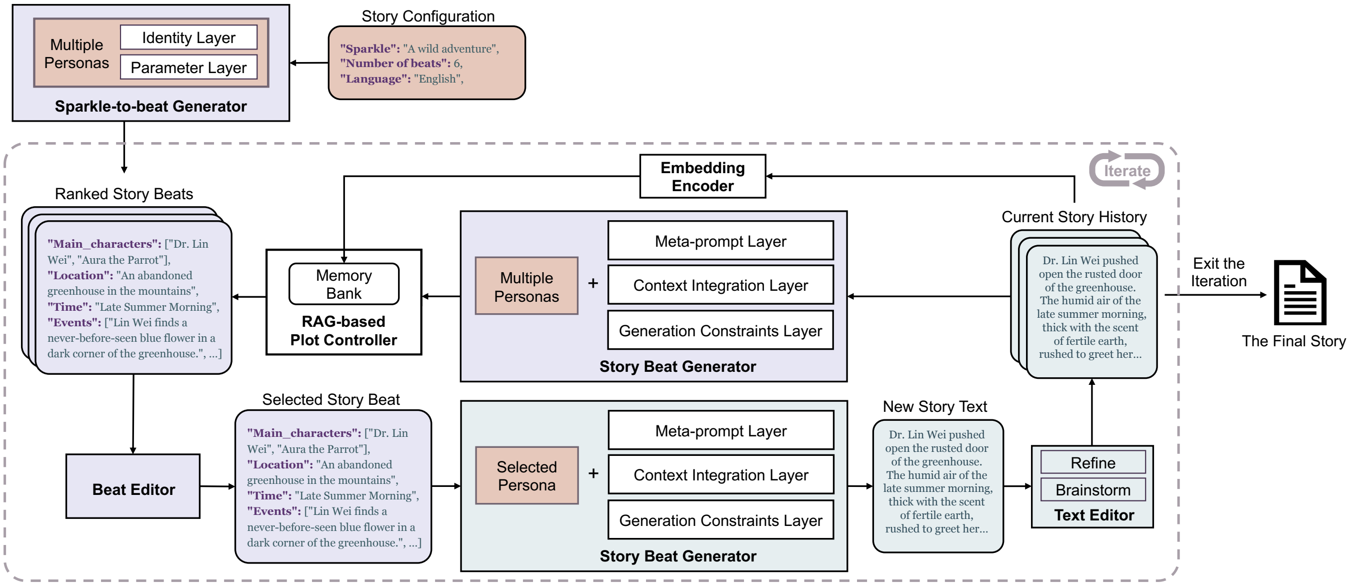


Figure 3: NarrativeLoom's technical pipeline. The system transforms user sparkles into story beats via multi-persona generation, employing a three-layer prompt architecture (meta-prompt, context integration, generation constraints) across both beat and text generation stages. The RAG-based Plot Controller ensures narrative consistency while users iterate through beat selection and text refinement. Purple indicates beat-level operations and green indicates text-level operations. Rectangles represent functional modules, while rounded corners represent data.

`key_events` for 3–5 pivotal actions. This structured format enables consistent processing while remaining interpretable during user selection. The system defaults to a six-beat structure following Hauge [30]'s framework, though users can configure 1–10 beats. Beat complexity adapts to narrative position: initial beats contain 3–4 events for world-building, while climactic beats expand to 4–5 events for dramatic intensity. Users can further adjust complexity through interactive refinement.

Stage-Specific Generation Logic. Sparkle-to-beat generation transforms the user's initial narrative seed into the first story beat, while subsequent beat generation builds upon established story elements.

Text Expansion Process. When the user selects a beat, the system expands it into 800–1000 words of narrative text. The expansion uses the same three-layer prompt architecture, with the selected persona passed as a parameter. This parameterized approach avoids the need for multiple persona-specific generators, reducing both cost and latency. Throughout generations, the system maintains up to 8000 tokens of narrative history to preserve context continuity.

4.3.2 Persona Instantiation and Parallel Generation.

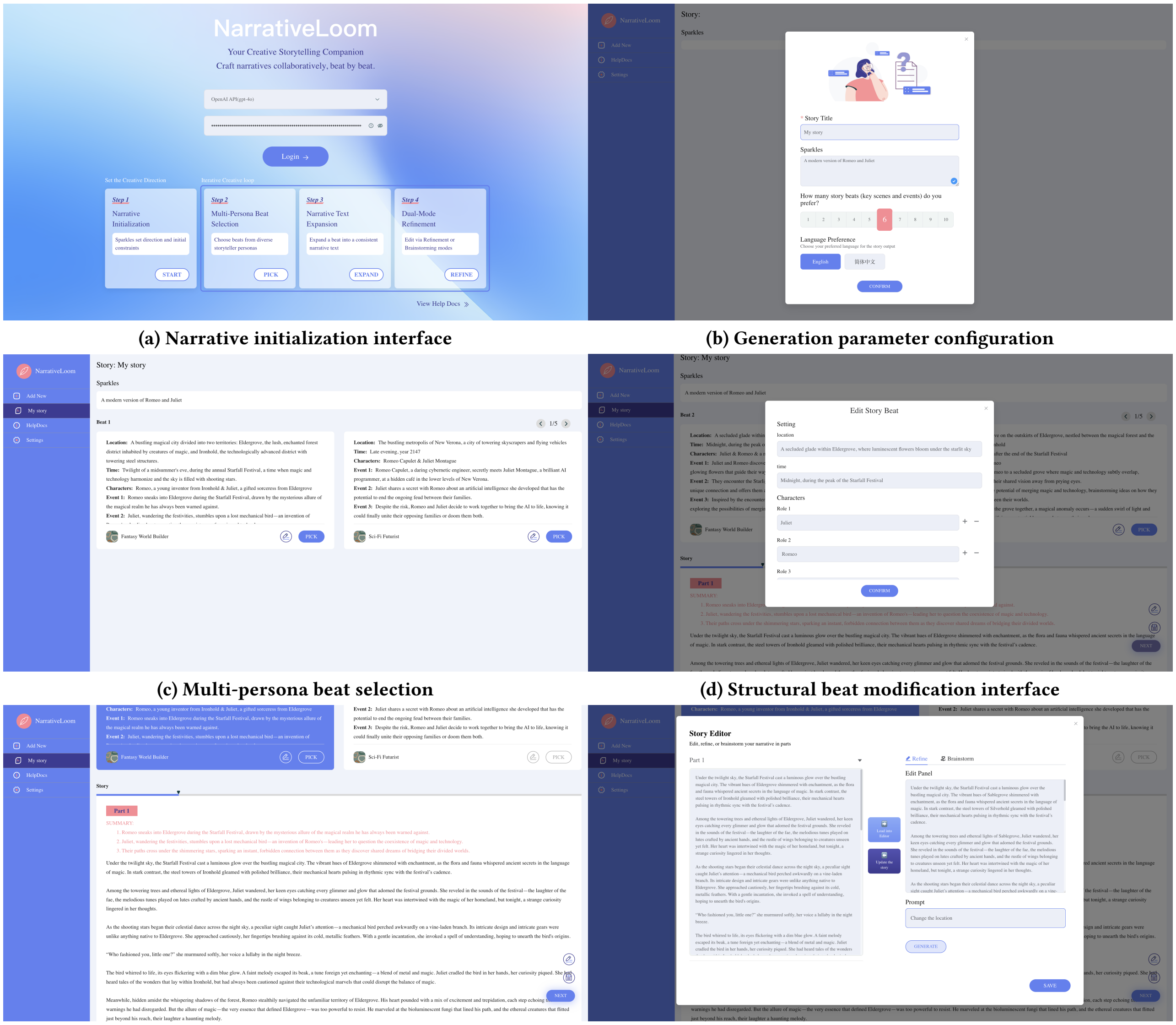
Prompt-Based Persona Instantiation. Each persona is defined through a multi-layered prompt template. The **identity layer** establishes genre-specific writing philosophies and creative priorities, while the **parameter layer** defines quantitative constraints on narrative elements (e.g., lexical diversity, dialogue-to-narrative ratios). Rather than providing static content instructions, this structure shapes the persona's generative behavior at a deeper level. Each persona is instantiated as a separate call to GPT-4o, with these layers driving consistent persona-specific variation across generations.

Parallel API Coordination. Rather than sequential generation, all ten personas generate story beats simultaneously through parallel API calls. This parallel architecture ensures that each persona produces independent alternatives without influence from other outputs. Each API call includes the complete story context (up to 8000 tokens of narrative history) and persona-specific instructions (approximately 500 tokens of role-defining prompts).

4.3.3 Interface Architecture and Interaction Mechanisms.

- **Brainstorm Module:** This module facilitates open-ended exploration by allowing users to interact with the AI to generate narrative inspiration. It serves as a conversational partner to help users overcome creative blocks, explore alternative plot developments, or elaborate on character motivations without directly modifying the current draft.
- **Refine Module:** This module provides direct editorial support by allowing users to submit specific revision requests. Users can instruct the AI to rewrite existing prose to adjust pacing, enhance descriptive detail, or modify dialogue for better character consistency. This ensures that the final text reflects the user's specific stylistic and narrative intent through iterative revision.

Interface Overview. The interface is built using Streamlit and employs a three-panel layout for multi-persona narrative exploration (see Fig. 4). The left sidebar manages story portfolios, the central workspace displays persona-generated content for comparison, and the lower panel supports narrative refinement through dual-mode editing. In the central workspace, each persona's beat proposal appears in a dedicated panel showing its role-specific rationale and templated content. The system maintains active session data in memory while persisting user story portfolios to storage.



(a) Narrative initialization interface

(b) Generation parameter configuration

(c) Multi-persona beat selection

(d) Structural beat modification interface

(e) Narrative text expansion

(f) Dual-mode refinement mechanism

Figure 4: The NarrativeLoom interface implementing the BVSR framework. The six panels illustrate the system's functional components: (a) narrative initialization interface; (b) generation parameter configuration; (c) multi-persona beat selection; (d) structural beat modification interface; (e) narrative text expansion; and (f) dual-mode refinement mechanism. This workflow supports iterative variation and selection while preserving human creative agency.

Beat Selection Display. The central workspace ranks beat proposals by consistency using the RAG-based plot controller (detailed in Sec. 4.3.4). The interface displays consistent options first and keeps all ten persona outputs accessible for exploration.

Editing and Refinement. The system supports editing at two levels to ensure human-AI alignment throughout the creative process. **Beat-level editing** allows users to modify settings, characters, and events before text expansion, with these structural changes propagating through the generation pipeline. **Text-level editing** enables prose refinement through manual intervention or two specialized AI-powered modules:

4.3.4 RAG-Based Plot Controller Implementation.

Technical Implementation. The plot controller maintains narrative coherence through a consistency check system utilizing Retrieval-Augmented Generation (RAG) via LlamaIndex [40]. The system detects logical inconsistencies between newly generated beats and established story elements by combining semantic retrieval with language model reasoning.

Story History Indexing. Story history is indexed via a vector-based memory mechanism. Each completed story segment is treated as a standalone document and encoded using the OpenAI `text-embedding-ada-002` model. This process generates 1536-dimensional embeddings that represent the semantic content of previous narrative developments. These embeddings are stored in a LlamaIndex vector store, allowing the system to retrieve relevant historical context during the evaluation of subsequent story beats.

Consistency Check Process. For each new story beat generated by a persona, the system executes contextual retrieval and logical verification. The beat content is converted into a query, embedded, and compared against the vector store using cosine similarity:

$$\text{sim}(\mathbf{v}_{\text{beat}}, \mathbf{v}_i) = \frac{\mathbf{v}_{\text{beat}} \cdot \mathbf{v}_i}{\|\mathbf{v}_{\text{beat}}\| \|\mathbf{v}_i\|},$$

where $\mathbf{v}_{\text{beat}}$ is the embedding of the verification query and $\mathbf{v}_i$ represents the i-th story history embedding. The retrieved contexts are provided to GPT-3.5-turbo with the following prompt: *"The new story beat is: [beat content]. Are there any logical errors in the events of the new story beat? Answer briefly in [Yes] or [No]. If [Yes], briefly describe the errors."* The LLM evaluates the beat for contradictions regarding prior events, character status (*e.g.*, unexplained revivals), temporal sequences, or established world rules. If an error is identified, the system does not discard the beat but labels the inconsistency for the user.

5 User Study

We conducted a within-subjects user study to evaluate the effectiveness of NarrativeLoom in supporting collaborative storytelling. Our evaluation compared the proposed system against a conventional single-persona chatbot interface to investigate its influence on the creative process, narrative quality, and user engagement. We hypothesized that the structured guidance and diverse narrative options provided by NarrativeLoom would enhance storytelling capabilities relative to the baseline chatbot.

5.1 Participants

Participants were recruited via Prolific, with eligibility restricted to native English speakers over 18 years of age holding at least an undergraduate degree. This criterion was established to ensure sufficient proficiency for complex narrative tasks. Of the 109 initial recruits, 94 passed a mandatory familiarization test, which served as a procedural fact-check to verify comprehension of the interface mechanics and core narrative rules.

To maintain high data quality, participants were excluded if they spent less than 20 minutes interacting with the systems or completed fewer than two story beats in the NarrativeLoom condition. This filtering process resulted in a final sample of **50 participants** for analysis. The final cohort (24 female, 26 male) had a mean age of 34.82 years (SD = 10.03, range: 22–71). Reported ethnicities included White (n=35), Asian (n=7), Black (n=6), Mixed (n=1), and Non-disclosure (n=1).

Diverse educational backgrounds were represented, including Arts and Humanities (n=11), Business (n=11), Natural Science (n=10), Social Science (n=10), and other fields (n=8). Regarding degree attainment, 33 participants held undergraduate degrees, 15 held graduate degrees, and 2 held doctoral degrees. Compensation was provided at a rate of £9/hour. The study protocol was approved by the university's Institutional Review Board (IRB).

5.2 Study Procedure and Protocol

We employed a within-subjects design in which participants interacted with both the experimental system (NarrativeLoom) and a baseline system. To mitigate order effects, system presentation was counterbalanced across participants [49]. The study followed five sequential phases:

1. **Demographics and Background (10 min):** Participants reported demographic data, writing experience, and genre preferences.
2. **Creative Warm-up (5 min):** Participants drafted initial story ideas ("sparkles") and wrote a short story for two minutes to establish a baseline for unaided creation.
3. **Familiarization and Training (10 min):** Participants reviewed an infographic of the protocol and completed two comprehension questions to verify task understanding.
4. **System Interaction (40 min):** Participants used each system for 20 minutes following a 3-minute tutorial per system.
   - **NarrativeLoom Condition (20 min):** Participants used our multi-persona system (GPT-4o API). For each narrative round, the system generated multiple continuations via anonymized AI personas. Users selected, edited, or built upon these options without knowing the specific persona source. Participants were required to complete at least two story beats using their sparkles.
   - **Chatbot Condition (20 min):** As a control, participants interacted with a custom chatbot interface powered by the same model using the same sparkle. Stories were developed through direct conversational prompting with a single AI agent.
5. **Post-Task Evaluation (10 min):** After each condition, participants completed a survey regarding their experience with the systems.

5.3 Data Collection and Analysis

We collected quantitative and qualitative data to evaluate the creative product, creative process, and narrative quality.

5.3.1 Creative Product and Process. We measured the quality of the creative product and process using metrics adapted from the Creative Product Semantic Scale (CPSS) [46] and the Creativity Support Index (CSI) [9, 14]. These adaptations were necessary because the original instruments were designed for different creative domains—CPSS for general creative products and CSI for broader creativity support tools—requiring domain-specific modifications for storytelling evaluation.

Our adaptations involved three key modifications:

(1) **Domain Specialization:** We translated abstract creativity dimensions into storytelling-specific constructs. For instance, CPSS's general "resolution" dimension was operationalized as narrative "coherence" (absence of plot inconsistencies), while CSI's "exploration" was reframed as narrative "diversity" (range of story possibilities).
(2) **Scale Simplification:** We streamlined the original 71-item CPSS and 12-item CSI into six focused dimensions using 5-point

Table 3: Expert evaluator demographics and professional experience. Profiles of the four expert reviewers, including their occupation, education, and years of professional writing experience.

ID	Gender	Age	Occupation	Education	Writing Experience (Years)
E1	Female	32	Screenwriter	Master of Fine Arts in Film Production	12
E2	Male	34	Screenwriter and Director	Master of Fine Arts in Film Production	8
E3	Female	38	Creative Producer	Bachelor of Arts in Chinese Literature	18
E4	Male	31	Journalist and Writer	Bachelor of Arts in Comparative Literature	10

Table 4: Torrance Test for Creative Writing (TTCW) evaluation framework. A systematic assessment consisting of 14 binary criteria across four creative dimensions (Fluency, Flexibility, Originality, and Elaboration) to evaluate narrative quality.

Dimension	Test Questions
Fluency	Does the manipulation of time feel appropriate and balanced? Does the story display awareness of balance between scene and summary? Does the story make sophisticated use of metaphor or literary devices? Does the end feel natural and earned, rather than arbitrary? Do story elements work together to form a unified whole?
Flexibility	Does the story provide diverse, convincing perspectives? Does the story balance interiority and exteriority effectively? Does the story contain turns that are both surprising and appropriate?
Originality	Will readers obtain unique ideas from this story? Is the story original without clichés? Does the story show innovation in structure or format?
Elaboration	Does the writer make the fictional world believable at the sensory level? Are characters developed at appropriate complexity levels? Does the story operate at multiple levels of meaning?

Likert scales, reducing participant fatigue while maintaining construct validity for comparative evaluation.

(3) **Contextual Relevance:** We reformulated questions to reflect AI-assisted storytelling contexts. For example, CSI's "collaboration" factor was adapted to "customization," measuring the system's ability to align AI-generated content with users' creative vision rather than human-human collaboration.

Participants rated their experience across six dimensions organized into two categories:

- **Product Metrics:**
  - *Novelty* (adapted from CPSS's "originality"): the degree of surprise and originality in generated narratives
  - *Diversity* (adapted from CSI's "exploration"): the range and variety of narrative possibilities offered
  - *Coherence* (adapted from CPSS's "resolution"): logical consistency and absence of plot contradictions
- **Process Metrics:**
  - *Customization* (adapted from CSI's "collaboration"): the system's responsiveness to user preferences and creative direction
  - *Engagement* (adapted from CSI's "enjoyment"): sustained interest and involvement during story creation
  - *Usability* (adapted from CSI's "results worth effort"): ease of interaction relative to creative output quality

5.3.2 Narrative Quality Analysis. To complement subjective ratings, we conducted computational text analysis of the generated stories using spaCy [32] and NLTK [6]. We analyzed four properties:

- **Word Count:** A basic metric of narrative length and development.
- **Gunning Fog Index:** A measure of text readability and complexity.
- **Dialogue Ratio:** The proportion of text presented as character dialogue *vs.* narrative exposition.
- **Location Count:** The number of unique spatial references, indicating the richness of the setting.

5.3.3 Persona Selection Tracking. For the NarrativeLoom condition, we tracked which AI personas users selected throughout the story creation process. This allowed us to analyze patterns in persona preference, frequency, and selection sequences.

5.4 Expert Review and Feedback

To validate our findings, we recruited four experts with extensive creative writing experience to evaluate the narrative quality of the generated stories (see Tab. 3 for demographics). The experts worked in two pairs (E1-E2, E3-E4). Each pair independently evaluated the same set of 10 story pairs (20 stories total) randomly sampled from our user study dataset, resulting in a total of 20 pairs and 40 stories evaluated across both expert pairs. To mitigate bias, all stories were anonymized, and their presentation order was randomized.

For the evaluation, we used the TTCW protocol [10], which provides a systematic assessment across four creativity dimensions through 14 binary tests (see Tab. 4). Each expert independently provided a binary pass or fail assessment for each test, accompanied by a brief justification. A story's final creativity score is the total number of tests passed (0–14). Following the TTCW evaluation, experts made a forced-choice comparison by selecting what they judged to be the superior story in each pair.

Finally, we conducted semi-structured interviews with all experts to gather qualitative insights on the perceived differences between the systems. The interviews explored five key areas: (i) noticeable differences in creative quality between systems, (ii) TTCW dimensions showing the largest quality gaps, (iii) patterns distinguishing higher- *vs.* lower-quality stories, (iv) recurring creative failures or limitations, and (v) priorities for improving AI storytelling systems. We recorded and thematically analyzed the interviews to identify common patterns across expert perspectives.

6 Findings

6.1 Strategic Exploration of a Diverse Narrative Landscape

To evaluate whether NarrativeLoom successfully enhanced creative exploration through diverse narrative possibilities, we analyzed both user perceptions of creativity and the underlying patterns of how users strategically engaged with the multi-persona system.

User-Perceived Creative Enhancement. In terms of user-perceived creative enhancement, NarrativeLoom demonstrated a practically meaningful advantage over the baseline (see Fig. 5). Analysis using a paired samples t-test confirmed that the improvement in **diversity** (see Sec. 5.3.1) was statistically significant ($M = 4.08$, $SD = 0.89$ *vs.* $M = 3.66$, $SD = 1.23$; $t(49) = 2.14$, $p = 0.037$, Cohen's $d = 0.39$). For **novelty** (see Sec. 5.3.1), while the higher ratings ($M = 4.20$, $SD = 0.98$ *vs.* $M = 3.94$, $SD = 1.10$; Cohen's $d = 0.25$) did not reach statistical significance, the effect size still indicated a trend favoring our system.

Users consistently praised NarrativeLoom's creative diversity, with P23, P24, and P34 noting it offered "*much more choice of plot direction*" and generated ideas that "*seemed more diverse and novel.*" Users particularly valued the structured approach to creative exploration, with P9 explaining that NarrativeLoom *"allows having the different beats added layers to the story instead of, in a chatbot, creating just one part of the story."* .

Frequency and Narrative Roles. An analysis of persona usage patterns revealed that users strategically selected specific personas for specialized roles depending on the narrative stage (see Fig. 6a). While the Historical and Dystopian personas were the most frequent choices for story initiation, the Mystery persona maintained the highest overall usage frequency. This contrast suggests an intuitive functional assignment by users: personas with strong world-building archetypes (Historical and Dystopian) served as "initiators," establishing the setting and initial conflict, whereas the Mystery persona emerged as the primary "developer," utilized to navigate plot complexities and advance the narrative once the story was underway.

Transitions and Narrative Flow. To understand how stories evolved, we analyzed the transition network between personas (see Fig. 6b). The transitions were not random but formed a structured network characterized by high-frequency pathways and strong directional preferences. The flow from Mystery to Romance was the most common, while other prominent paths included Adventure to Comedy and Dystopian to Horror. These pathways often represent complementary genre pairings, suggesting users were actively blending genres to create richer narratives. Furthermore, this flow was significantly asymmetric, with the average difference between forward and reverse transitions being statistically significant ($M = 1.24 \pm 0.92$, $t(45) = 8.95$, $p < 0.001$). For example, the transition from Mystery to Romance (5 instances) was more than twice as common as the reverse (2 instances), while the Historical to Magical path (3 instances) was never reciprocated. This directionality demonstrates that users guided their stories along specific trajectories, leveraging the multi-persona system not for simple variety, but as a toolkit for purposeful, structured narrative escalation.

6.2 Balancing Co-Creative Agency with Scaffolding

Quantitatively, NarrativeLoom performed comparably to the baseline across three dimensions (see Sec. 5.3.1): **customization** (NarrativeLoom: $M = 4.10, SD = 0.92$ *vs.* chatbot: $M = 3.98, SD = 0.97$), **usability** (NarrativeLoom: $M = 4.42, SD = 0.83$ *vs.* chatbot: $M = 4.36, SD = 0.84$), and **engagement** (NarrativeLoom: $M = 4.00, SD = 1.15$ *vs.* chatbot: $M = 3.96, SD = 1.04$), with no statistically significant differences. This suggests that managing multiple personas did not compromise usability despite the added complexity. Beyond these quantitative similarities, our qualitative analysis reveals distinct forms of creative control enabled by NarrativeLoom. We identify three key dimensions of enhanced co-agency.

Distributed Creative Authority (Shared Agency). NarrativeLoom's multi-persona architecture distributed creative authority between system and user. Rather than producing a single linear suggestion, the personas generated diverse narrative directions that expanded the creative space. Users remained in charge of selecting, combining, or discarding these options, thereby maintaining clear editorial control. Participants valued this balance of shared agency; for instance, P23 remarked, "*I liked the different options from the agents*," while P24 emphasized that it offered "*much more choice of plot direction*" compared to a traditional chatbot that typically steers the story along a single path.

Responsive Collaborative Refinement (Negotiated Agency). Users experienced agency as a negotiated process where their feedback actively redirected the system's output. Multiple participants emphasized the system's ability to integrate their feedback into subsequent generations. For example, P19 noted that it "*took into account edits and comments alongside coming up with engaging scenarios and characters.*" This iterative responsiveness illustrates how user input and system creativity co-shaped the evolving trajectory of the story.

Empowered Creative Partnership (Augmented Agency). By amplifying users' expressive capacities, NarrativeLoom shifted the experience toward an empowered sense of agency. Participants described how the system "*generated a comprehensive and creative story using my prompts (P30)*" and produced narratives that "*read so*

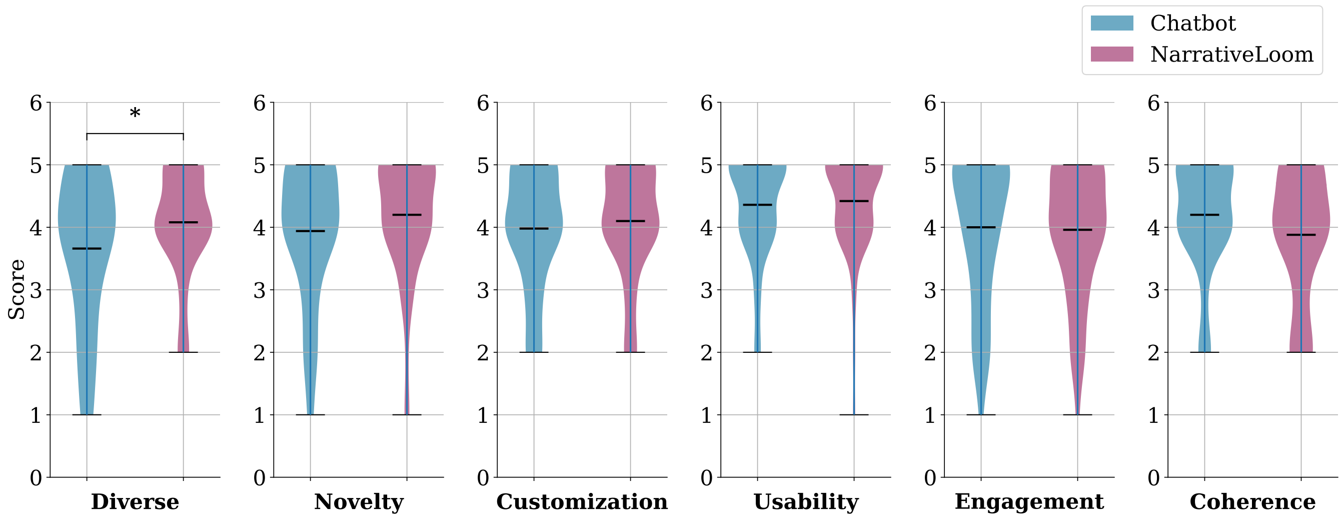
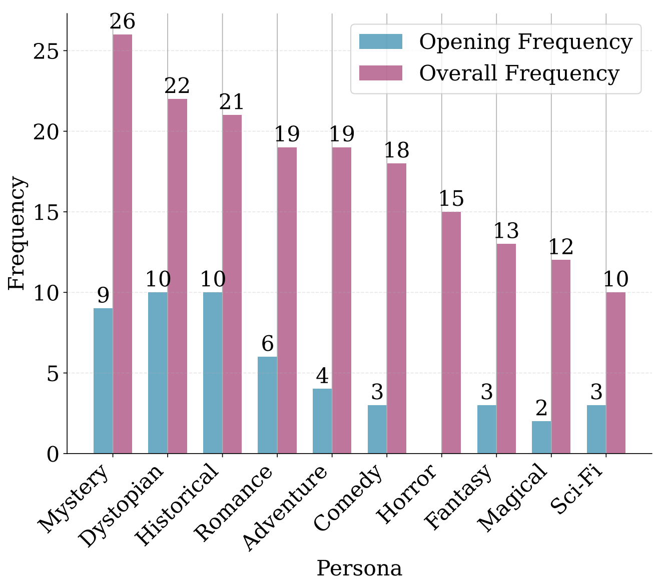
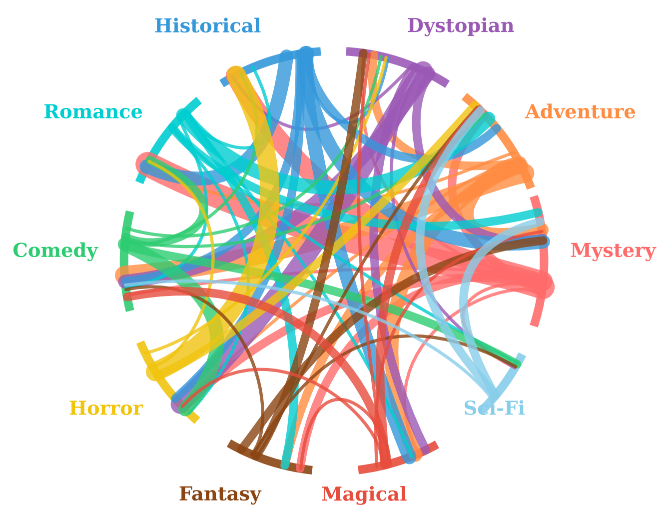


Figure 5: The comparison of user evaluation scores between NarrativeLoom and the chatbot. Violin plots show the distribution of user ratings on a 5-point scale across key dimensions. NarrativeLoom consistently achieved higher median scores for creativity-focused metrics like diversity and novelty, while performing comparably on usability, engagement, and coherence. Asterisks denote statistical significance: * indicates $p < 0.05$.



(a) Persona usage roles. The grouped bar chart reveals specialized roles. Personas like Historical and Dystopian were primary "initiators" with high opening usage, while Mystery was the key "developer," dominating overall usage.



(b) Narrative flow network. The transition graph shows that movement between personas was not random. Edge thickness indicates frequency, revealing popular, often asymmetric pathways (*e.g.*, from Mystery to Romance).

Figure 6: Users strategically selected personas and transitioned between them in structured patterns. The data reveals two key user behaviors: (a) selecting personas for specialized "initiator" or "developer" roles, and (b) moving between personas along logical, genre-adjacent pathways.

much more naturally and fuller compared to the others (P24)." P45 reflected this heightened agency through strong engagement: "*I love the different stories the system generated for me. It was captivating and really interesting.*"

6.3 Navigating the Creativity-Coherence Trade-off with a Segmented Structure

NarrativeLoom's structured approach, which breaks the narrative into manageable creative units, yielded significant advantages in story development and richness (see Fig. 7). NarrativeLoom enabled users to create substantially **longer** stories ($M = 3803.16, SD = 1109.50$ words) than the chatbot baseline ($M = 1907.88, SD = 1304.44$ words; $t(49) = 9.160, p < .001$). This structure also supported the creation of richer narrative environments, with stories from NarrativeLoom incorporating significantly more **locations** ($M = 3.86, SD = 2.67$ *vs.* $M = 2.44, SD = 2.54$; $t(49) = 3.276, p = .002$) and a higher ratio of **dialogue** ($M = 0.30, SD = 0.12$ *vs.* $M = 0.16, SD = 0.13$; $t(49) = 5.675, p < .001$). Furthermore, the resulting narratives were more accessible, achieving a lower (better) **readability** score ($M = 7.45, SD = 0.54$ *vs.* $M = 8.17, SD = 1.16$; $t(49) = -4.212, p < .001$).

The qualitative feedback confirmed the value of this segmented process. As P10 noted, "I liked how NarrativeLoom approached the storytelling process in sections. chatbot felt more all or none while the system worked on one moment at a time." However, this same multi-persona, segmented architecture that fostered creativity also introduced the challenge of maintaining narrative coherence.

Indeed, the chatbot baseline achieved a higher **coherence** rating (see Fig. 5; $M = 4.20, SD = 0.92$) compared to NarrativeLoom ($M = 3.88, SD = 1.05$).While this difference was not statistically significant, it represented a small-to-medium effect size (Cohen's $d = 0.32$). This is an expected trade-off: the chatbot generates a continuous, single-voice narrative, while our system's multi-persona design and significantly longer stories create more potential fragmentation points where consistency can be challenged.

Despite these results, users reported that NarrativeLoom maintained logical consistency. P28 noted that "overall the suggestions and generated text were good and the process of generating the story was smooth," while P38 observed that the system "created compelling chunks that followed through to create a fluid story." These qualitative reports suggest that the system manages the coherence-creativity trade-off, increasing narrative richness while maintaining acceptable story coherence.

6.4 Writing Expertise Shapes Preferences for AI Writing Assistance

User characteristics, specifically writing experience, influenced interactions with the storytelling systems, suggesting a need for personalized AI writing assistance. We observed different patterns in how novice and senior writers evaluated the two systems (see Fig. 8), although the interaction effects were not statistically significant.

Regarding novelty, novice writers rated NarrativeLoom higher than the chatbot baseline (NarrativeLoom: $M = 4.26, SD = 0.90$; chatbot: $M = 3.78, SD = 1.18$; Cohen's $d = 0.457$). Conversely, senior writers showed no clear preference, rating both systems similarly (chatbot: $M = 4.00, SD = 1.00$; NarrativeLoom: $M = 3.94, SD = 1.18$).

Usability ratings followed a different pattern. Novices reported a slight preference for NarrativeLoom (NarrativeLoom: $M = 4.39, SD = 0.97$; chatbot: $M = 4.26, SD = 0.90$; Cohen's $d = 0.140$), whereas senior writers rated the chatbot's usability higher (chatbot: $M = 4.67, SD = 0.47$; NarrativeLoom: $M = 4.39, SD = 0.76$; Cohen's $d = 0.441$).

User preferences were also influenced by their specific writing stage. P8 summarized this distinction: *"Chatbot is suitable for one who knows their rough story already and just needs help making it. But the NarrativeLoom system for those who want to write but are not sure what story to tell."* This suggests that AI assistance should align with different phases of the creative process. The structured BVSR framework in NarrativeLoom supports the ideation and exploration phase by helping writers identify narrative possibilities. In contrast, traditional chatbots may be more effective during the development phase, when writers focus on executing and refining established concepts.

These results indicate that a universal approach to AI writing assistance may be suboptimal. The observed preference patterns—novices prioritizing the novelty of NarrativeLoom and senior writers prioritizing chatbot usability—suggest that support systems should be tailored to both user expertise and their current stage in the writing process.

6.5 Expert Evaluation Results

To complement the large-scale user study, four creative writing experts assessed narrative quality. We randomly sampled 20 story pairs from the dataset, each consisting of one story generated by NarrativeLoom and one by the chatbot baseline ($N = 40$). The experts demonstrated a near-unanimous preference for NarrativeLoom. While noting that neither system consistently produced perfect outputs, the experts rated the NarrativeLoom narratives as higher quality and more compelling than those from the baseline.

Quantitative Preference and Creativity Scores. Experts preferred NarrativeLoom in direct comparisons, selecting stories generated by NarrativeLoom in 38 out of 40 forced-choice scenarios. This preference is reflected in the TTCW scores (Fig. 9). A paired t-test showed that NarrativeLoom achieved higher overall creativity scores ($M = 9.72$ out of 14, $SD = 2.1$) than the chatbot baseline ($M = 5.00$ out of 14, $SD = 1.8$; $t(79) = 13.68, p < 0.001$). Analysis of individual dimensions showed advantages across fluency ($M = 4.38$ out of 5 *vs.* 2.40; $t(79) = 12.93, p < 0.001$), flexibility ($M = 2.33$ out of 3 *vs.* 1.15; $t(79) = 12.61, p < 0.001$), elaboration ($M = 1.95$ out of 3 *vs.* 1.07; $t(79) = 8.31, p < 0.001$), and originality ($M = 1.07$ out of 3 *vs.* 0.38; $t(79) = 6.16, p = 0.001$). Furthermore, NarrativeLoom outperformed the chatbot baseline across all user expertise levels; the magnitude of this advantage did not differ significantly between novice and senior writers (Fig. 10). Although the small expert sample size limits broad statistical inference, these results align with our larger-scale user study.

Creative Originality and Unpredictability. Expert qualitative feedback identified three primary areas where NarrativeLoom produced superior narratives, centering first on creative innovation. The baseline was frequently criticized for relying on convention; E1 noted that it *"frequently relies on conventional story structures and clichéd elements."* In contrast, NarrativeLoom was lauded for its narrative surprise. As E4 remarked, *"These stories take you to a place that you don't expect... The chatbot stories are kind of predictable."* This was echoed by E3, who noted, *"This story generated by NarrativeLoom is a bit of a surprise. I didn't expect it when reading the [sparkle]."* Experts also pointed to specific novel ideas, with E1 highlighting

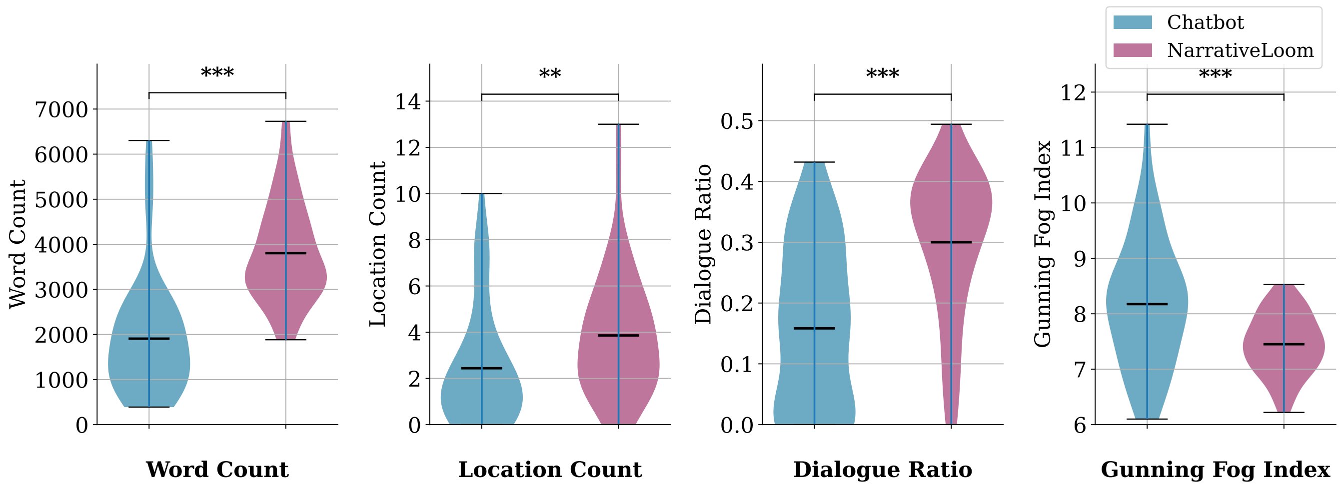


Figure 7: The comparison of narrative features between NarrativeLoom and the chatbot. Violin plots compare the chatbot baseline (blue) with NarrativeLoom (pink). NarrativeLoom generated stories that were significantly longer, contained more locations, had a higher ratio of dialogue, and were more readable (lower Gunning Fog Index). Asterisks denote statistical significance: ** indicates $p < 0.01$ and * indicates $p < 0.001$.**

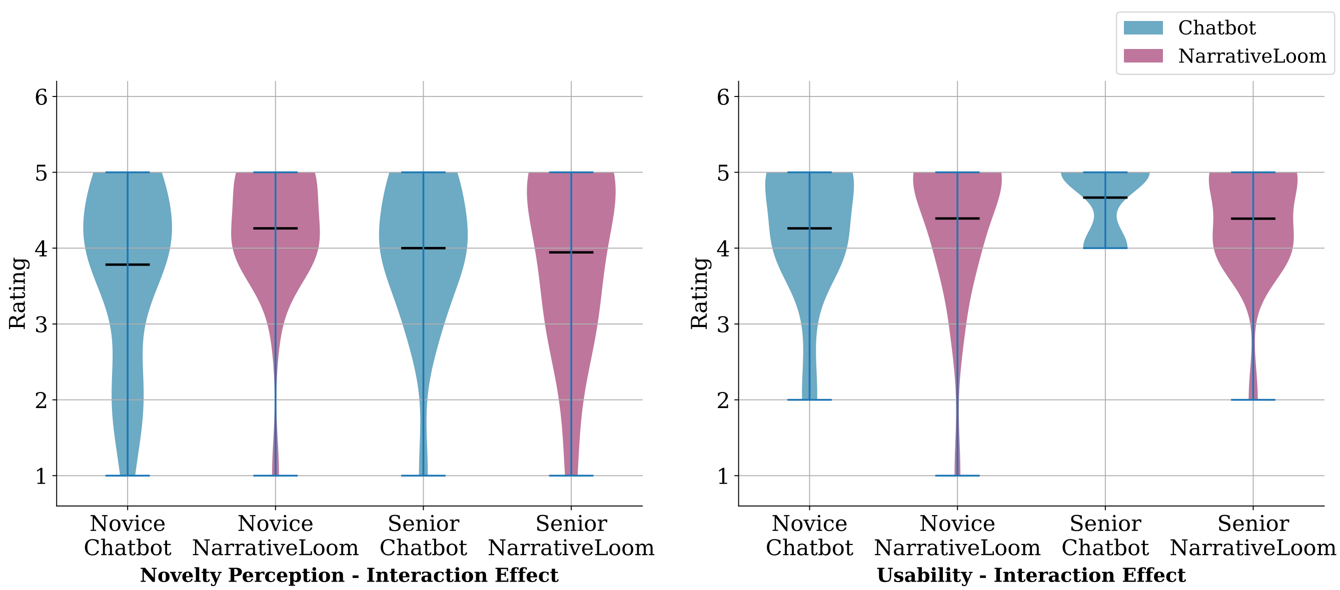


Figure 8: Interaction between writing expertise and system preference. Violin plots show user ratings for novelty perception (left) and usability (right), comparing NarrativeLoom (purple) against the chatbot baseline (blue). Novice writers perceived NarrativeLoom as more novel, whereas senior writers rated the chatbot higher on usability.

that *"stories from NarrativeLoom often have some clever ideas or highlights, things that I haven't seen before and are less clichéd. For example, for the 'friendship' topic, a time traveler appears after the friends meet."*

Superior Narrative Craft: "Showing" Over "Telling". A second theme involved NarrativeLoom's use of immersive, scene-based storytelling. E1 contrasted the two systems by describing that the chatbot *"is a bit like an instruction manual, laying things out directly from a 'God's eye view' and simply listing events."* In contrast, E1 observed that *"NarrativeLoom doesn't present all the information at once; instead, it starts with a scene, giving the feeling of a story being slowly and engagingly told."* This aligns with E2's comment that NarrativeLoom's writing is *"deeply internal, rich with details and expanded scenic development,"* whereas the chatbot is *"more third-person, objective, and summary-like."* The distinction was articulated most directly by E4: *"NarrativeLoom has a much higher level in the way it is written. It is better at showing not telling."*

Character Depth and Psychological Realism. Finally, experts noted NarrativeLoom's ability to render characters with psychological depth, a quality they found lacking in the baseline. E1 observed that

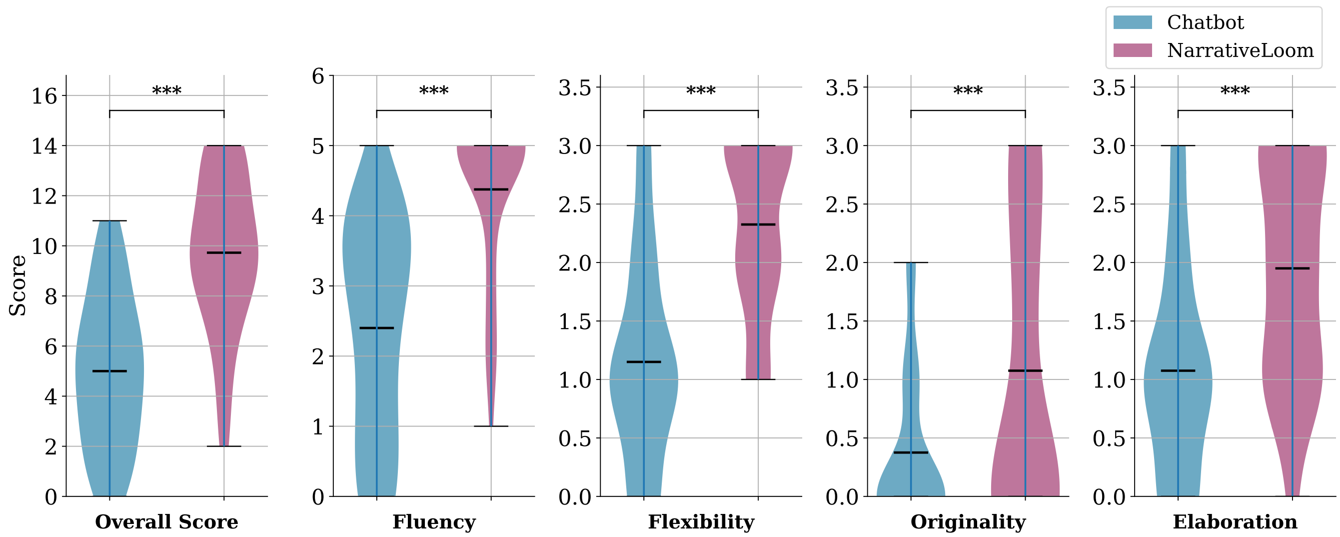


Figure 9: Expert evaluation results using the TTCW framework. Violin plots show the distribution of creativity scores across four dimensions and overall performance, comparing NarrativeLoom (purple) against the chatbot baseline (blue). NarrativeLoom demonstrated statistically significant improvements across all creativity dimensions. Asterisks denote statistical significance: * indicates $p < 0.001$.**

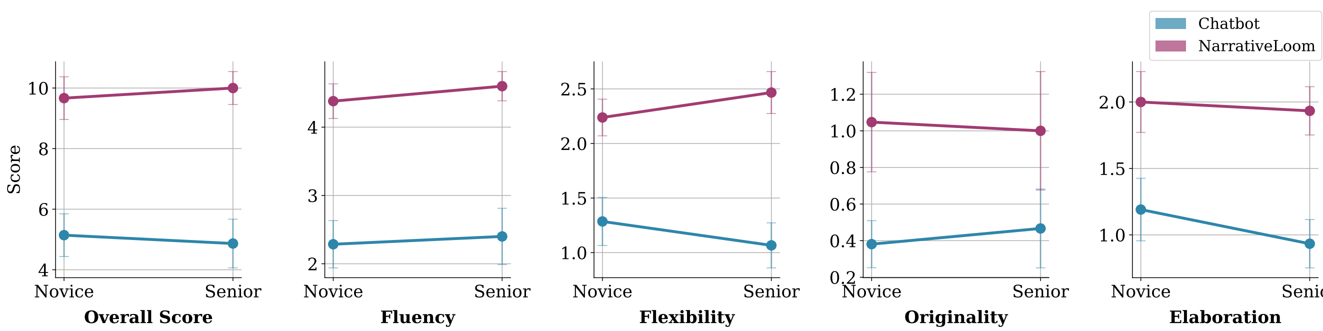


Figure 10: Expert TTCW evaluations participant narratives based on writing expertise. Mean scores ($\pm SE$) are shown for novice and senior writers using NarrativeLoom (purple) and chatbot (blue). Evaluation criteria include Overall Score, Fluency, Flexibility, Originality, and Elaboration. NarrativeLoom outperformed the chatbot across all metrics for both user groups.

"most characters in NarrativeLoom's stories are well set," and E4 summarizing the preference directly by stating that NarrativeLoom *"has better characters and dialogues."* E2 elaborated on this distinction, noting that chatbot characters *"lacked internal depth and psychological complexity with no inner hesitation or psychological processing."* Conversely, NarrativeLoom produced more nuanced figures; E2 highlighted NarrativeLoom's *"strong character development,"* particularly regarding antagonists who felt like *"real people rather than artificial constructs."* This capacity for believable interiority was a consistent differentiator across the expert reviews.

7 Discussion

Our study showed that NarrativeLoom significantly improved creative storytelling compared to the chatbot system (see Sec. 6.5), with users strategically leveraging different personas for specialized narrative roles (see Sec. 6.1), producing longer, richer-detailed stories (see Sec. 6.3), and maintaining creative agency despite the system's structured guidance (see Sec. 6.2). These findings illuminate how Campbell's BVSRs theory [8] can serve as a principled framework for designing AI systems that enhance creative possibilities while preserving user agency.

7.1 Balancing Creative Diversity with Cognitive Manageability

Our findings reveal a fundamental tension in AI-assisted creative writing: how to expand narrative possibilities while maintaining cognitive manageability. Current systems address this through interventions at different narrative layers. Systems operating at the **discursive layer** through parametric sampling (Wordcraft [68]) enable stylistic variation but cannot alter underlying story logic. The **emotional trajectory layer** (TaleBrush [15]) provides visual story arc control but constrains narrative specificity. The **structural planning layer** offers either rigid top-down cascading (Dramatron [43]) that limits spontaneity or configurable abstraction (WhatElse [41]) that demands complex pre-planning.

Our genre-based persona approach intervenes at the **diegetic layer**—the level of story events and causal logic. This distinction matters because genres shape narrative possibility spaces [13]. By instantiating multiple genre personas, we address the statistical centroid problem [5], where single models converge on predictable outputs by sampling from the center of their training distribution. Each persona samples from distinct narrative regions: when participants used comedy for tension relief or mystery for plot complexity, they accessed fundamentally different narrative logics. This explains why experts found NarrativeLoom stories more likely to take readers to unexpected places while baseline stories remained predictable—genre intervention alters *what can happen*, not just *how it's told*.

The beat-based design provides a specific granularity for creative scaffolding. Dramatron's [43] rigid planning ensures coherence but demands heavy cognitive pre-planning. Wordcraft's [68] open-ended generation offers flexibility but can overwhelm users. Our approach provides a middle-out position—semantic anchors that enable structured improvisation without predetermined outcomes. This operates within the zone of proximal development [62]: it is complex enough for sophisticated narratives yet modular enough to remain manageable. The significantly longer stories, richer dialogue, and diverse settings in NarrativeLoom validate this balance. Minor coherence trade-offs represent acceptable costs for enhanced creative exploration.

7.2 Temporal Synthesis and Collaborative Emergence

The locus of creativity in BVSR remains a subject of ongoing debate. While classical theory identifies variation as the primary creative driver—Campbell [8] and Simonton [55, 56] argue that creativity arises from numerous blind variations—recent computational research suggests that creative agency resides in human selection. In this view, internal predictive models guide strategic choices rather than stochastic filtering [20, 72].

Our results indicate that creativity in NarrativeLoom emerges from a triadic interaction between algorithmic variation, human selection, and contextual synthesis. Persona transition patterns (see Fig. 6b) show that users do not merely select options; they construct narrative trajectories through sequential choices, where each selection constrains subsequent possibilities. Asymmetric transitions suggest that users exploit narrative affordances [61] arising from the intersection of multiple personas. This observation aligns with Zhou and Lee [70], who noted in text-to-image contexts that human creativity manifests through curation. It further supports Epstein et al. [22]'s argument that generative AI shifts creativity from direct manipulation to iterative specification, allowing users to maintain control over the narrative trajectory.

However, our findings identify a distinction relevant to HCI design: unlike image generation [70], which involves selecting from parallel alternatives, narrative creation requires temporal synthesis. Users must integrate selected beats into coherent sequential structures. This temporal dimension facilitates collaborative emergence [52], where creative products arise from structured improvisation between human and AI. The higher dialogue ratios and increased location diversity in NarrativeLoom stories suggest that this synthesis amplifies creative elaboration beyond individual capabilities, supporting theories of distributed cognition [34]. These results suggest that co-creative systems should function as platforms for temporal synthesis rather than simple generation tools, leveraging both human curation and AI variation.

7.3 Writers Require Varying Degrees of Creative Support

Our findings reveal that writers require different types of creative support based on their **expertise level** and **creative stage** (see Sec. 6.4), challenging the one-size-fits-all approach in current AI writing tools. This aligns with prior work suggesting that writers seek diverse roles from AI collaborators [11, 27], yet illuminates a critical design consideration often overlooked in creativity support research [14].

Consistent with expert-novice differences in creative domains [23, 24], we observed separate preference patterns. Novice writers favored NarrativeLoom's structured exploration, benefiting from multi-persona scaffolding that activates different narrative spaces [57] and helps overcome creative blocks when their own resources are limited [65, 69]. Conversely, expert writers preferred streamlined, chatbot-like interfaces that integrate seamlessly into their established workflows. This aligns with the expertise reversal effect [36]—scaffolding beneficial to novices becomes extraneous cognitive load for experts who possess highly developed schemas and seek to maintain creative flow [17].

Crucially, our TTCW evaluation revealed that these preference differences did not translate into quality differences—both groups achieved comparable creative improvements with NarrativeLoom. This paradox illuminates a key distinction: creative support operates at two levels—**workflow integration** (where experts favor minimal disruption) and **cognitive stimulation** (where structured diversity benefits all). The multi-persona system's value lies in combating cognitive fixation through systematic variation, a benefit that persists regardless of expertise level.

Beyond static expertise, support needs vary dynamically by creative phase. Writers may benefit from NarrativeLoom's exploratory scaffolding during ideation but prefer streamlined assistance during revision [12, 25, 35]. This suggests future creative AI should implement **adaptive dual-channel designs**: preserving cognitive benefits of structured variation while allowing interface customization from fully scaffolded (novices/ideation) to minimally intrusive (experts/refinement). The key principle is decoupling creative scaffolding from interface complexity—maintaining the former's value while minimizing the latter's disruption.

7.4 Design Implications for Creative AI Systems

7.4.1 Theory-Guided Design as a Counterpoint to Model Scale. Our results show that creative AI systems benefit from deliberate, theory-guided frameworks rather than relying solely on increased parameter scale to improve performance [7, 48]. The success of our structured multi-persona model, grounded in Campbell's BVSR framework [8], suggests that principled design can be more effective than raw computational power. This aligns with arguments that current AI approaches may face diminishing returns without fundamental structural innovation [42]. Instead of pursuing larger models, our approach demonstrates that implementing diverse creative perspectives allows systems to explore narrative spaces often underrepresented in training corpora, such as those requiring novel genre combinations or unconventional plot developments. Future creative AI could employ modular systems with specialized components optimized for specific creative functions, providing more targeted assistance.

7.4.2 Structured Improvisation Architecture for Creative Emergence. The beat-based segmentation approach addresses a core challenge in long-form generation: maintaining coherence without stifling emergence. Unlike hierarchical planning systems that may constrain discovery [43] or unconstrained generation that risks incoherence [31], our model enables progressive coherence—local consistency within manageable units paired with emergent global development. These findings suggest that creative AI should decompose complex tasks into discrete, context-aware units that preserve spontaneity while providing a navigational structure for extended development [53]. Future systems could use adaptive mechanisms to automatically identify optimal boundaries based on project type, user preference, and context, scaling from short-form to extended works while maintaining both novelty and logic.

7.4.3 Explicit Role Separation for Creative Agency. The balance between AI assistance and human ownership in NarrativeLoom highlights the necessity of explicit role separation. By assigning AI the responsibility for variation generation and reserving selection authority for the human, our approach maintained user agency while leveraging computational speed. Creative AI systems should therefore implement transparent boundaries between machine and human contributions, avoiding black box patterns that undermine creator confidence [18, 37, 71]. Future designs should include explicit role indicators, separate interfaces for generation and selection, and allow users to negotiate these boundaries by choosing how much creative territory to delegate.

7.5 Limitations and Future Work

While our results demonstrate the effectiveness of BVSR-based computational creativity, several theoretical and methodological limitations warrant consideration. The reliance on genre-based persona specialization, while effective for narrative diversity, may tend to draw upon established literary conventions rather than fully exploring unconventional creative possibilities [13]. Our evaluation framework, though comprehensive, necessarily reflects Western narrative traditions and assessment criteria. Cross-cultural validation would be essential to establish broader applicability, particularly given evidence that creativity manifestation and evaluation vary significantly across cultural contexts [45]. Additionally, the observed expertise effects raise questions about the long-term developmental implications of AI-assisted creativity. While novices benefited from system support, the long-term developmental impact remains unclear. Longitudinal studies are necessary to determine whether such scaffolding facilitates skill acquisition or inadvertently creates dependency, potentially impeding the development of independent creative capabilities. This concern mirrors established debates regarding technological scaffolding in educational contexts [47], where over-reliance on external support can diminish the internal cognitive effort required for mastery. Investigating this trajectory is vital for designing systems that serve as developmental tools rather than mere cognitive crutches.

Future work will prioritize four system enhancements. First, we will **extend beyond genre-based specialization through function-oriented modules**, developing personas focused on world-building, character psychology, and narrative pacing. This hybrid system aims to enable unconventional creative combinations while retaining the genre diversity benefits observed in our study. Second, we plan to **implement adaptive weighting mechanisms** that learn from user selection patterns to adjust persona prominence. By proactively suggesting complementary alternatives whenever the system senses a decline in creative divergence, this approach seeks to mitigate creative fixation. Third, we will **incorporate culturally-aware variants** trained on literary traditions beyond Western narratives. This requires collaboration with cultural experts to ensure authentic representation of story structures and conventions across global contexts [45]. Finally, we will **develop expertise-sensitive interfaces** that dynamically adjust scaffolding based on user proficiency. The system will track selection speed, edit frequency, and coherence scores to provide exploration support for novices and streamlined tools for experts [51], adapting to both user capability and task demands.

8 Conclusion

The successful implementation of BVSR theory in NarrativeLoom demonstrates that psychological creativity frameworks can provide effective blueprints for computational creativity systems. By explicitly separating variation generation from selective retention and implementing structured diversity through specialized personas, we have created a creative partnership model that enhances human creativity while preserving creative agency. The expertise-dependent preferences observed in our study highlight the importance of adaptive design in creative AI, while the strategic persona utilization patterns suggest that humans can effectively collaborate with AI ensembles when provided with appropriate interaction frameworks.

These results suggest that the next frontier of creative AI lies not solely in smarter models, but in principled design that respects the nuances of individual creative cognition. As AI capabilities continue to evolve, theoretically-informed approaches will be vital in ensuring that AI remains a partner in the creative process—fostering systems that genuinely extend the reach of human imagination while safeguarding the fundamental human drive for expression and ownership.

Acknowledgments

We extend our gratitude to Zhen Chen for creating the beautiful illustrations that enhance this paper. We also thank Yujia Peng, Guangyuan Jiang, and Yizhou Wang for their valuable feedback and insightful suggestions throughout the development of this work. This work is supported in part by the National Science and Technology Innovation 2030 Major Program (2025ZD0219400), the National Natural Science Foundation of China (62376009, 62376031), the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei

Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: a survey of approaches. *Comput. Surveys* 54, 5 (2021), 1–38.

[3] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Conference on Creativity & Cognition*.

[4] Aristotle. 1942. *The poetics of Aristotle*. University of North Carolina Press.

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *ACM Conference on Fairness, Accountability, and Transparency*.

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Donald T Campbell. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review* 67, 6 (1960), 380.

[9] Erin A Carroll and Celine Latulipe. 2009. The creativity support index. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[10] 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 *ACM Conference on Human Factors in Computing Systems (CHI)*.

[11] Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 1–33.

[12] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Conference on Creativity & Cognition*.

[13] Daniel Chandler. 1997. An introduction to genre theory.

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

[15] 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 *ACM Conference on Human Factors in Computing Systems (CHI)*.

[16] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *International Conference on Intelligent User Interfaces*.

[17] Mihaly Csikszentmihalyi et al. 1997. Flow and the psychology of discovery and invention. *HarperPerennial, New York* 39 (1997), 1–16.

[18] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An enactive model of creativity for computational collaboration and co-creation. *Creativity in the Digital Age* (2015), 109–133.

[19] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-ai collaboration: varied scaffolding levels in co-writing with language models. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[20] Arne Dietrich and Hilde Haider. 2015. Human creativity, evolutionary algorithms, and predictive representations: The mechanics of thought trials. *Psychonomic Bulletin & Review* 22, 4 (2015), 897–915.

[21] Steven Earnshaw. 2014. *Handbook of Creative Writing*. Edinburgh University Press.

[22] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[23] K Anders Ericsson. 2018. The Differential Influence of Experience, Practice, and Deliberate Practice on the Development of Superior Individual Performance of Experts. In *The Cambridge Handbook of Expertise and Expert Performance*, K Anders Ericsson, Robert R Hoffman, Aaron Kozbelt, and A Mark Williams (Eds.). Cambridge University Press, 745–769.

[24] K Anders Ericsson and Jacqui Smith. 1991. *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press.

[25] Ronald A Finke, Thomas B Ward, and Steven M Smith. 1996. *Creative cognition: Theory, research, and applications*. MIT press.

[26] Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & Society* (2024), 1–11.

[27] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[28] Pablo Gervás. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30, 3 (2009), 49–49.

[29] Hans Hansen, Daved Barry, David M Boje, and Mary Jo Hatch. 2007. Truth or consequences: An improvised collective story construction. *Journal of Management Inquiry* 16, 2 (2007), 112–126.

[30] Michael Hauge. 2011. *Writing screenplays that sell*. Bloomsbury Publishing.

[31] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[32] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://github.com/explosion/spaCy.

[33] David Howard and Edward Mabley. 1993. *The tools of screenwriting: A writer's guide to the craft and elements of a screenplay*. Macmillan.

[34] Edwin Hutchins. 2000. Distributed cognition. *International Encyclopedia of the Social and Behavioral Sciences* 138, 1 (2000), 1–10.

[35] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030* (2022).

[36] Slava Kalyuga. 2009. The expertise reversal effect. In *Managing Cognitive Load in Adaptive Multimedia Learning*. IGI Global Scientific Publishing, 58–80.

[37] Anna Kantosalo and Hannu Toivonen. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *International Conference on Computational Creativity*.

[38] Joy Kim, Justin Cheng, and Michael S Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Conference on Computer Supported Cooperative Work & Social Computing*.

[39] Randee Lipson Lawrence and Dennis Swiftdeer Paige. 2016. What our ancestors knew: Teaching and learning through storytelling. *New Directions for Adult and Continuing Education* 149, Spring (2016), 63–72.

[40] Jerry Liu. 2022. *LlamaIndex*. https://doi.org/10.5281/zenodo.1234

[41] Zhuoran Lu, Qian Zhou, and Yi Wang. 2025. WhatELSE: Shaping narrative spaces at configurable level of abstraction for AI-bridged interactive storytelling. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[42] Gary Marcus. 2020. The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177* (2020).

[43] 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 *ACM Conference on Human Factors in Computing Systems (CHI)*.

[44] Michael Z Newman. 2006. From beats to arcs: Toward a poetics of television narrative. *The Velvet Light Trap* 58, 1 (2006), 16–28.

[45] Weihua Niu and Robert J Sternberg. 2001. Cultural influences on artistic creativity and its evaluation. *International Journal of Psychology* 36, 4 (2001), 225–241.

[46] Karen O'Quin and Susan P Besemer. 1989. The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal* 2, 4 (1989), 267–278.

[47] Roy D. Pea. 2004. The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. *Journal of the Learning Sciences* 13, 3 (2004), 423–451.

[48] Yongqian Peng, Yuxi Ma, Mengmeng Wang, Yuxuan Wang, Yizhou Wang, Chi Zhang, Yixin Zhu, and Zilong Zheng. 2025. Probing and Inducing Combinational Creativity in Vision-Language Models. In *Annual Meeting of the Cognitive Science Society (CogSci)*.

[49] William D Perreault. 1975. Controlling order-effect bias. *The Public Opinion Quarterly* 39, 4 (1975), 544–551.

[50] Mark O Riedl and R Michael Young. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24 (2006), 303–323.

[51] Gavriel Salomon, David N Perkins, and Tamar Globerson. 1991. Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher* 20, 3 (1991), 2–9.

[52] R Keith Sawyer. 2000. Improvisation and the creative process: Dewey, Collingwood, and the aesthetics of spontaneity. *The Journal of Aesthetics and Art Criticism* 58, 2 (2000), 149–161.

[53] R Keith Sawyer. 2014. *Group creativity: Music, theater, collaboration*. Psychology Press.

[54] R Keith Sawyer and Stacy DeZutter. 2009. Distributed creativity: How collective creations emerge from collaboration. *Psychology of Aesthetics, Creativity, and the*

Arts 3, 2 (2009), 81.

[55] Dean Keith Simonton. 1999. Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry* (1999), 309–328.

[56] Dean Keith Simonton. 2023. The blind-variation and selective-retention theory of creativity: Recent developments and current status of BVSR. *Creativity Research Journal* 35, 3 (2023), 304–323.

[57] Ut Na Sio, Kenneth Kotovsky, and Jonathan Cagan. 2015. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies* 39 (2015), 70–99.

[58] Andrea J Stone. 1995. *Images from the underworld: Naj Tunich and the tradition of Maya cave painting*. University of Texas Press.

[59] Theresa Jean Tanenbaum and Karen Tanenbaum. 2008. Improvisation and performance as models for interacting with stories. In *International Conference on Interactive Digital Storytelling*.

[60] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[61] Brad R Turner. 2025. Narrative Affordances: What Stories Can and Cannot Do. *Academy of Management Annals* ja (2025), annals–2023.

[62] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard University Press.

[63] Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F Karlsson. 2023. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomputing* (2023), 126792.

[64] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[65] Robert W Weisberg. 2006. Modes of Expertise in Creative Thinking: Evidence from Case Studies. In *The Cambridge Handbook of Expertise and Expert Performance*, K Anders Ericsson, Neil Charness, Paul J Feltovich, and Robert R Hoffman (Eds.). Cambridge University Press, 761–787.

[66] Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving Long Story Coherence With Detailed Outline Control. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[67] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

[68] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *International Conference on Intelligent User Interfaces*.

[69] Huan Yuan, Kelong Lu, Mengsi Jing, Cuirong Yang, and Ning Hao. 2022. Examples in creative exhaustion: The role of example features and individual differences in creativity. *Personality and Individual Differences* 189 (2022), 111473.

[70] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (2024), pgae052. https://doi.org/10.1093/pnasnexus/pgae052

[71] Jianlong Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *IEEE Conference on Computational Intelligence and Games*.

[72] Yuxi Zhu, Simone M Ritter, Barbara CN Müller, and Ap Dijksterhuis. 2017. Creativity: Intuitive processing outperforms deliberative processing in creative idea selection. *Journal of Experimental Social Psychology* 73 (2017), 180–188.