Plotania: Exploring Transparency Trade-offs in AI Co-Writing Through Virtual Readers and Transparent Attribution

Yufeng Hu
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
huyf23@mails.tsinghua.edu.cn

Jinyi Zhang
RMIT University
Melbourne, Victoria, Australia
chelsea082400@gmail.com

Zehuan Wang
School of Creative Media
City University of Hong Kong
Hong Kong, China
wangzehuan2023@outlook.com

Chun Yu*
Ministry of Education
Key Laboratory of Pervasive Computing
Beijing, China
Beijing Key Laboratory of Networked Multimedia
Beijing, China
Tsinghua University
College of AI
Beijing, China
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

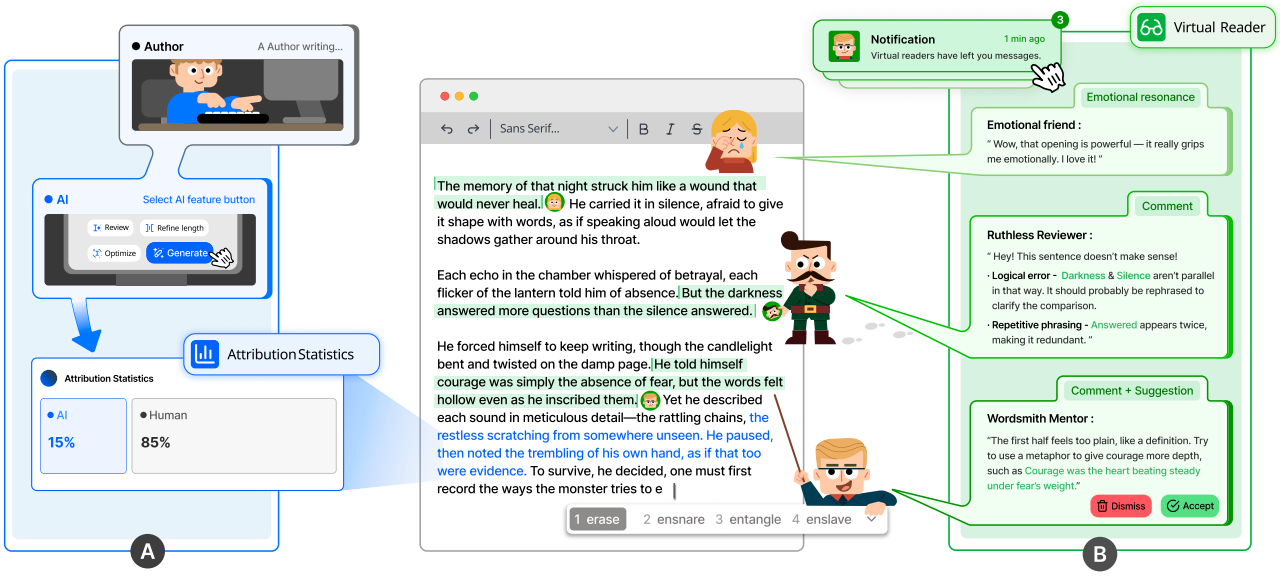

Figure 1: Plotania provides two key capabilities: real-time attribution tracking and virtual reader feedback. Interface shows (A) real-time transparent attribution statistics, and (B) virtual reader providing real-time feedback.

Abstract

Current AI writing tools aim to enhance authorial capacity yet often diminish authorial control and lack timely audience feedback.

*Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790926

Through a formative study with fiction authors (N=10), we uncovered two critical tensions in human–AI co-writing: balancing AI scaffolding with authorial ownership, and the absence of contextual audience perspectives that shape storytelling during drafting. Guided by these insights, we designed Plotania, a co-writing system that combines proactive virtual readers offering real-time audience reactions with transparent attribution layers. A controlled study (N=20) revealed *complex and counterintuitive effects*: virtual reader feedback increased audience awareness but *decreased* perceived creative agency, transforming individual authorship into collaborative performance. Transparent attribution raised awareness of AI contributions but triggered identity anxiety and reduced AI usage. These findings reveal fundamental trade-offs in transparency design. We contribute design principles for "agency-preserving transparency" that balance information provision with creative empowerment, informing future transparency design in human-AI creative collaboration.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**; *Interactive systems and tools*; • **Computing methodologies** → *Natural language generation.*

Keywords

creative agency, narrative control, audience-aware feedback, human-AI co-writing

ACM Reference Format:

Yufeng Hu, Jinyi Zhang, Zehuan Wang, and Chun Yu. 2026. Plotania: Exploring Transparency Trade-offs in AI Co-Writing Through Virtual Readers and Transparent Attribution. 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, 23 pages. https://doi.org/10.1145/3772318.3790926

1 Introduction

Large language models promise to accelerate fiction writing, yet they disrupt the intimate dialogue between author and reader that has defined storytelling for millennia. Writers navigating AI assistance face fundamental tensions—not just around authorial control, but around the very audience awareness that shapes compelling narratives.

The dialogue between author and imagined audience forms the foundation of fiction writing [10, 66], embodying what literary theorists describe as the triadic nature of narrative creation: the author (subject) engages in writing (action) directed toward an audience (object) [6, 34]. Writers craft narratives by continuously anticipating reader responses. This audience-centered process exemplifies what we term creative agency—**the writer's perceived individual control over creative decisions and authorial ownership of the narrative** [12, 69]. Creative agency in AI-assisted writing involves two interdependent processes: understanding one's authorial contributions ("who I am") and maintaining audience awareness ("for whom I write").

Current AI co-writing systems, while offering substantial benefits, disrupt both ends of this triadic relationship. On the *author side*, **opaque attribution** prevents writers from distinguishing their contributions from AI-generated content [40, 49], blurring creative boundaries and diminishing the authorial identity essential to creative ownership [47]. On the *audience side*, **absent audience perspectives** limit systems to generic suggestions divorced from specific reader contexts [21, 76], severing the "audience anchor" that guides narrative decisions and leaving writers without the situated feedback that shapes compelling storytelling.

Existing research has made progress through ideation support [21], variation exploration [63], and scaffolding mechanisms [27]. Attribution methods have emerged to track AI contributions [25], and audience feedback systems have shown value in interactive contexts [55]. However, these approaches treat transparency and audience awareness as separate design concerns rather than interdependent processes. This separation misses a crucial insight: addressing one dimension without the other provides incomplete support for creative agency. Attribution transparency alone enables writers to know "what I wrote" but offers no directional purpose for "why I wrote this way"; audience feedback provides external validation but may obscure individual contributions. Moreover, these mechanisms could generate tensions reflecting enduring literary dialectics between individual authorship and social performance [13], intrinsic motivation and external evaluation [3]. Understanding whether these mechanisms synergistically enhance agency, conflict to diminish it, or produce complex conditional effects requires empirical investigation. No existing work has examined how they interact when integrated.

To understand how writers navigate these challenges and identify design opportunities, we ask:

- **RQ1.** What tensions arise when writers use AI assistance, and how do they negotiate creative agency?
- **RQ2.** How do writers maintain authorial ownership while benefiting from AI collaboration?
- **RQ3.** What are the effects of transparent attribution and audience-aware feedback on perceived creative agency in co-writing systems?

A controlled study with twenty hobbyist and emerging writers reveals a fundamental transparency paradox: mechanisms designed to enhance human agency can paradoxically diminish it. Despite high engagement with virtual reader feedback, participants experienced complex and counterintuitive agency effects. Reader feedback *decreased* perceived creative agency by transforming individual authorship into collaborative performance management, while transparent attribution created "algorithmic anxiety" that reduced AI usage as writers avoided assistance to maintain favorable contribution scores. However, when combined, these mechanisms demonstrated complementary rather than simply additive effects, with modest improvements suggesting that transparency in creative contexts requires careful orchestration rather than maximization. Most unexpectedly, participants developed *parasocial relationships* with AI personas, seeking emotional companionship that challenged traditional boundaries between tools and creative partners, revealing the complex psychological landscape of human-AI creative collaboration.

Contributions.

- **Creative Agency Framework.** Empirical insights into how writers negotiate creative agency in human-AI collaboration,

revealing specific strategies writers develop to maintain authorial control amid transparency tensions.
- **System Contribution.** Plotania, a co-writing system integrating virtual readers and transparent attribution layers to investigate trade-offs between information provision and creative agency preservation in AI collaboration.
- **Complex Effects and Design Principles.** Controlled evaluation (N=20) revealing that transparency mechanisms produce complex and counterintuitive effects on creative agency, contributing design principles for "agency-preserving transparency" that balance information with empowerment rather than maximizing visibility.

2 Related Work

Fiction writers have traditionally exercised creative control over their narratives, transforming personal visions into distinctive voices. AI writing systems now enable direct collaboration between authors and language models, raising questions about maintaining authorial identity and creative control in these partnerships. Two aspects are particularly relevant to co-writing design: understanding AI contributions and their impact on authorial ownership, and integrating audience perspectives that have traditionally guided narrative craft.

2.1 Creative Agency and Attribution in Human-AI Writing

Creative agency, which refers to the writer's perceived control over creative decisions and authorial ownership [12, 69], becomes complex in human-AI collaboration. Unlike traditional writing where agency resides solely with the author, human-AI co-writing distributes creative control across human and model, with initiative, authorship, and accountability shifting dynamically [47, 48]. This challenge intersects with longstanding theoretical questions about authorship [9, 11, 35].

Agency Threats in Attribution. Research identifies three threats to authorial agency: *boundary invisibility*, *role-boundary conflicts*, and *accountability ambiguity* [16, 47, 54, 71]. These challenges are compounded by users attributing more creativity to AI as they grant it more autonomy [40].

Transparency Solutions and Limitations. Research addresses attribution challenges through transparency, which involves making AI contributions visible [7, 25, 40, 67, 73]. While XAI techniques exist [1, 65], creative domains require approaches that preserve creative identity [14]. However, transparency can paradoxically reduce agency through multiple pathways. Disclosure of AI involvement may negatively impact writing quality evaluation [19], while transparency about AI capabilities may increase authors' overreliance on it rather than guiding reasonable use [72]. Furthermore, making AI contributions visible may trigger evaluation anxiety as writers monitor quantified contribution metrics, transforming creative processes into performance management. These paradoxical effects suggest transparency solutions assume more information improves agency preservation, yet psychological effects on writers' creative experience remain underexplored [48].

Empirical Strategies for Preserving Agency. Writers selectively use AI as a brainstorming partner while restricting it in voice-critical passages [39, 44]. Value alignment between human and AI affects perceived ownership [38], while clear decision rights strengthen creative self-efficacy [52, 68]. Studies reveal distinct collaboration patterns varying by writing phase [58] and domain-specific concerns about creative control [36, 70]. Large-scale analysis of in-the-wild writing sessions confirms these patterns, revealing that users engage in diverse collaboration behaviors including revising intents, exploring alternative outputs, and iteratively refining text through under-specified multi-turn feedback [56].

Design Responses. Recent systems attempt to preserve agency through compositional substrates [17, 50] and by mapping AI strategies onto writing phases [27, 64]. These approaches build on established human-AI interaction guidelines [4] and computational creativity frameworks [22, 59]. However, they primarily focus on workflow orchestration rather than addressing transparent attribution. While attribution methods exist in other AI domains [25], writing tools lack real-time mechanisms that help authors track their contributions.

Inherent Tensions in Creative Authorship. Beyond technological solutions, creative agency exists within inherent tensions that have long characterized literary production. Writers navigate competing demands between individual artistic expression and audience expectations [13], intrinsic creative motivation and external evaluation pressures [3], and writer-based versus reader-based composition [33, 34]. Creative writing research identifies the fundamental tension between self-determination and external goals [8, 30]: expected evaluation constitutes one of the most detrimental extrinsic constraints on creative work [3], while maintaining autonomy proves essential for creative self-efficacy [24, 30]. The cognitive shift from writer-centered thought to audience-calibrated prose itself imposes significant cognitive demands [33]. AI-mediated authorship may transform these enduring tensions: transparent attribution quantifies creative contributions in ways that make evaluation salient, while computational audience feedback concretizes the abstract "imagined reader" into immediate judgment. Understanding how these mechanisms interact with foundational creative tensions remains unexplored, yet may prove crucial for designing agency-preserving collaborative writing systems.

2.2 Co-Writing Tools and the Missing Audience Perspective

While attribution challenges affect writers' sense of ownership, a second dimension of creative agency remains equally important yet less explored: the writer's ability to craft narratives through continuous anticipation of reader response. Current co-writing tools approach fiction writing as an individual creative process rather than recognizing it as an ongoing dialogue between author and imagined audience. This challenge connects to reader-response theory [32, 45, 66], though computational approaches to modeling reader response [42] remain underexplored in real-time writing tools.

Current Co-Writing System Approaches. Modern co-writing systems employ multiple design strategies: interaction granularity [37, 74], interface flexibility [5, 62], and workflow alignment [18, 27, 61, 64, 75]. These approaches focus on efficiency but do

not address authorship clarity or integrating audience perspectives that shape narrative decisions.

Design Challenges in Audience Integration. Fiction writing involves anticipating reader response [10], yet co-writing systems provide generic suggestions divorced from audience context [44, 76]. While other domains integrate real-time feedback [55] and recent work explores LLM-powered audience personas [20], writing tools face significant design challenges when integrating audience perspectives. Immediate feedback, despite being helpful, can disrupt creative flow and shift focus toward external validation rather than intrinsic exploration [31]. Moreover, audience feedback introduces power dynamics and normative pressures that may compromise creative autonomy, transforming individual authorship into social performance. Evidence from real-world usage reveals that writers generating professional documents frequently ask questions to learn domain norms and seek feedback on their writing [56], suggesting latent demand for audience-oriented guidance that current systems inadequately address. These challenges require carefully balancing computational assistance with creative autonomy [23, 26, 46, 53], particularly when feedback timing and framing may fundamentally alter the creative experience.

Contemporary co-writing systems have made significant advances in interaction design and workflow support. However, two aspects remain less thoroughly investigated: how transparency mechanisms in collaborative writing tools affect user experience and creative processes, and how writing systems might better support the audience-oriented nature of narrative craft. These gaps represent opportunities for further exploration in human-AI creative collaboration research.

3 Understanding Creative Agency in Writer-AI Collaboration

To understand how fiction writers navigate creative agency tensions in human-AI co-writing, we conducted a semi-structured study with 10 experienced fiction writers. We interviewed experienced writers because they can articulate nuanced creative practices and agency concerns that inform tools for our target users: hobbyist and emerging writers who may benefit most from transparency mechanisms as they develop their craft. This investigation directly addresses RQ1–RQ3 and provides foundational insights that inform Plotania's design principles.

3.1 Participants

We recruited 10 fiction writers through writing communities, writing forums, and targeted social media outreach. Our participants included 6 web-serialists, 2 fan-fiction authors, and 2 animation script writers, representing diverse writing contexts and audiences. Writing experience ranged from 1–10+ years, with most participants (6/10) having 4–6 years of experience.

Participants varied considerably in their AI usage patterns: 2 used AI occasionally (monthly), 7 frequently (weekly), and 1 heavily (daily). Common AI applications included ideation, outlining, polishing, and overcoming writer's block. Most participants actively serialize their work on platforms, which feature real-time reader feedback systems and highly engaged reader communities that significantly influence narrative development decisions. All interviews were conducted remotely with informed consent, audio recording, and $20 compensation. Detailed demographics appear in Appendix B.

3.2 Procedure

Each interview lasted 30–60 minutes and followed a structured timeline and artifact walkthrough methodology. Participants selected a recent writing project and screen-shared their notes, drafts, or planning documents while narrating their creative process across five stages: ideation, outlining, drafting, revision, and feedback incorporation.

We probed specific human-AI collaboration practices, asking participants to describe concrete LLM interactions including when and why they invoked AI assistance, their criteria for accepting or rejecting AI suggestions, and their strategies for maintaining creative control. To understand audience feedback preferences, we asked participants to identify one or two focal passages from their work and discuss their ideal feedback mechanisms for those sections.

The interviews concluded with a collaborative speculation session where participants envisioned their ideal co-writing and feedback tools. The complete interview guide is provided in Appendix A.

3.3 Analysis

We employed thematic analysis following established procedures. Two researchers independently coded all transcripts, achieving substantial inter-rater agreement ($\kappa = 0.82$) on 30% double-coded data. Data collection reached saturation at N=10. Chinese quotations were translated by native-speaking authors and back-checked for meaning preservation. Our analysis was structured around three core axial themes:

(1) **Creative workflow**: how writers capture and organize ideas and materials; generate, revise, and integrate across tools and devices.
(2) **Mindset and motivation**: goals, values, pain points, and perceived self-efficacy during creation.
(3) **Human-AI collaboration practices**: division of labor and boundaries, prompting strategies, trust and adoption, and experiential feedback.

3.4 Findings

From our analysis of these three axial dimensions, six cross-cutting themes emerged that reveal critical tensions in how writers maintain creative agency while collaborating with AI systems. These findings uncover not only established challenges around scaffolding and ownership, but also unexpected tensions around voice authenticity, audience calibration failures, and the paradoxical psychology of creative control.

3.4.1 F1. Flexible support without creative constraint [RQ1, RQ2]. Most participants (7/10) leveraged AI for **lightweight structural support**, generating outlines or beat lists to overcome initial creative blocks, while actively resisting rigid structures that might constrain their exploratory process. Writers consistently described a preferred workflow: "I set the topic, let AI generate an outline, then I complete it and later polish" (I8). However, they emphasized

that "the demanding work is turning the outline into actual prose; AI only gives a rough skeleton" (I8).

Participants expressed strong resistance to inflexible planning tools: "they force a fixed logic chain; I prefer plain text so I can change it anytime" (I9). While a minority preferred visual planning approaches (I5), the dominant pattern was clear: support tools are welcomed when they remain **mutable and negotiable**, but rejected when they calcify into rigid templates.

3.4.2 F2. Negotiating "authenticity": voice, initiative, and credit [RQ1, RQ2]. While writers appreciated AI support, this collaboration raised deeper questions about creative authenticity. Writers evaluated quality not merely through technical fluency, but through markers of personal ownership: continuity of voice, visibility of creative initiative, and confidence in authorship claims. A majority of participants (7/10) expressed concerns about AI-generated content diluting their authorial identity. As I5 explained: "If I use too much AI, it's no longer me writing—I'm not the first author, only a collaborator; it feels inauthentic, like there's a deceptive component." I9 echoed this sentiment: "Using AI directly feels like disrespecting creativity; it involves originality issues, and I can't grasp the proportion well." I7 similarly struggled with this tension: "I don't want the AI flavor to be too strong when AI and human collaborate, but I find it hard to control the level of AI involvement effectively."

These concerns translated into concrete requests for transparency mechanisms. I5 directly articulated the need: "It would be best if future writing tools could show me the AI-human ratio in my article at any time, so I can control the proportion through timely revisions." I10 described attempting to manage this through existing tools: "If the AI probability is high when writing, I hope to reduce it through plagiarism checking and then revise before submitting." Beyond quantitative attribution, writers' authenticity concerns also manifested as a desire for style sovereignty rather than generic AI assistance. I4 articulated this preference clearly: "I want a model of my own style, not one stuffed with everyone's styles." Participants also emphasized legal and professional ownership, with I10 noting that "the copyright is still mine."

Interestingly, comfort with AI delegation varied by expertise and creative stage. Some participants (e.g., I6) were willing to delegate high-level creative decisions like theme, character development, or tone when they felt less confident in those areas. This suggests that authenticity is contextually negotiated rather than absolute.

Beyond ownership concerns, multiple participants (4/10) noted a specific linguistic challenge: AI-generated text felt "official" or "mechanical," undermining personal voice. I4 explained: "AI's creativity is limited and lacks human touch...tends toward official writing," while I5 observed: "AI dialogue has logical connections, but humans don't actually speak that way." This "official language barrier" creates a distinctive threat to voice preservation, where AI's formal linguistic tendencies can systematically erode the informal, emotionally authentic tone that distinguishes personal creative expression. The central tension thus operates at multiple levels: efficiency gains versus ownership signals, and logical precision versus emotional authenticity.

3.4.3 F3. Genre-specific audience calibration failures [RQ1, RQ3]. The authenticity challenges described above are compounded by another fundamental limitation: AI systems fundamentally misunderstand genre-specific audience expectations, providing suggestions that contradict established reader communities and emotional tones. I7 noted: "AI doesn't understand the emotional tone of my web fiction—I write romance, it gives me patriotic plots." I9 observed that "The plots generated by AI seem mediocre and overly clichéd," suggesting AI draws from generic rather than genre-specific patterns.

These failures reveal a critical gap in current AI systems' audience awareness capabilities. While writers possess sophisticated understanding of their specific reader communities, AI lacks the contextual knowledge to provide genre-appropriate suggestions, forcing writers to invest substantial effort in correcting or filtering AI recommendations to match audience expectations.

3.4.4 F4. Audience-in-the-loop, but controllable [RQ1, RQ3]. Given AI's limitations in understanding audiences, nearly all participants (9/10) expressed strong desire for more sophisticated feedback mechanisms, specifically proposing that AI could simulate reader perspectives to provide targeted feedback. I6 articulated this vision: "AI is essentially a mapping of human thinking, so I think it can simulate reader thinking and give me various suggestions." I8 requested: "I hope AI can summarize after reading my article and tell me if there's anything that attracts readers or increases reading interest." I4 envisioned AI functioning "as a reviewer helping you see where there are flaws and giving polishing suggestions. Not just polishing, but also discussing whether the plot is stimulating enough, or where it's not deep enough or too shallow, giving suggestions that I can appropriately adopt."

Participants further specified a need for diverse reader personas with different critical stances. I6 explained: "AI's suggestions are relatively gentle...I need it to use sharper, more incisive language to comment or find faults, providing better guidance so I can identify problems in time," distinguishing between "ruthless critics" and "friend commenters." However, participants simultaneously sought to avoid the emotional disruption of unfiltered reader comments. I7 noted: "most authors avoid reading comments to protect their mindset."

This finding highlights a significant gap in current co-writing systems, which provide generic suggestions without contextual audience perspectives. Writers demonstrated sophisticated understanding of different feedback perspectives and their appropriate applications. I8 explained: "peers tell me whether they want to keep reading; teachers check if the logic is clear." This distinction guided how they sought and applied different types of feedback to their revision process.

A significant gap emerged in the availability of high-level structural feedback. I9 noted: "In typical online writing communities, you rarely get pacing or theme-level advice." Instead of seeking more feedback volume, participants wanted feedback that functioned like a targeted reader: bounded in both scope and critical stance, grounded in specific textual evidence rather than general impressions. This represents a desire for structured, constructive feedback rather than the emotionally disruptive noise of open comment systems.

3.4.5 F5. Granular delegation; long-context continuity [RQ1, RQ2]. While writers sought better audience feedback, they also articulated

clear preferences for how AI should assist in the actual writing process. Nearly all participants (9/10) described a remarkably consistent division of labor in their human-AI collaboration. Writers retained control over high-level creative elements (theme development, narrative arcs, plot twists, and voice) while viewing AI as well-suited for detailed micro-work including transitions, connective tissue, polishing, and consistency checking.

This delegation pattern reflects writers' strategic approach to preserving agency while leveraging AI efficiency gains. The pattern reflected both creative preferences and practical efficiency. I8 stated: "Transitions, connective passages, and basic polishing—I'm fine delegating entirely to AI." I9 reinforced this sentiment from a different angle: "filling so many transitions isn't joyful—it's painful." Writers saw these micro-tasks as necessary but creatively unfulfilling work that could be safely automated.

However, a critical technical limitation emerged around long-context coherence. I10 explained the frustration: "I want project-level memory; starting a new chat feels like resetting." Similarly, I6 noted that current models "struggle to absorb 1,000+ words at once," leading to repetitive content or narrative degradation over longer passages. This limitation directly impacts writers' ability to maintain narrative ownership across extended texts, as AI assistance becomes inconsistent with story-level coherence.

Writers thus articulated a clear need for granular delegation capabilities with appropriate guardrails, coupled with tools that maintain story-level memory and proactively surface rhythm or duplication issues before they compound into larger narrative problems.

3.5 Design Goals

Drawing directly from our five key findings (F1–F5), we derived five design goals that guide Plotania's development. These goals address three core challenges revealed by our analysis: preserving voice authenticity against AI's formal tendencies and attribution concerns (DG1, DG4 responding to F2), enabling genre-aware audience feedback while maintaining creative autonomy (DG2, DG3 responding to F3, F4), and providing flexible yet persistent collaborative support (DG1, DG5 responding to F1, F5).

3.5.1 DG1. Transparent authorship and edit lineage. To address writers' authenticity concerns (F2), ensure complete transparency in authorship attribution at all textual levels. Plotania provides granular authorship tracking (human vs. AI attributions) and maintains a chronological lineage of all insertions, deletions, and rewrites. The system provides reversible diffs at sentence, paragraph, and scene granularity, allowing writers to trace and undo any changes. When attribution becomes genuinely ambiguous (e.g., through heavy collaborative paraphrasing), the system explicitly marks uncertainty rather than asserting false precision. This approach ensures writers can maintain clear ownership while enabling precise control over AI contributions.

3.5.2 DG2. Genre-calibrated reader feedback. Addressing the genre-specific calibration failures (F4) and desire for controllable feedback (F5), replace unfiltered comment walls with controllable Virtual Readers that are explicitly calibrated to specific genre conventions and reader communities. These readers provide targeted feedback from specific perspectives (Romance Reader, Ruthless Reviewer, Plot Pace Master) understanding genre-unique emotional tones and narrative expectations. Writers control readers across scope (span/scene/arc), genre expertise, and temporal cadence (including quiet hours). Each comment anchors to specific textual evidence with genre-appropriate reasoning, ensuring feedback aligns with actual reader community expectations rather than generic advice.

3.5.3 DG3. Negotiable, lightweight structural support. Responding to writers' need for flexible support without creative constraint (F1), implement a **plot canvas** that writers can easily expand, contract, reorder, and replace without structural constraints. AI suggestions appear as **negotiable drafts** rather than fixed templates, preserving agency to modify or reject structures. The system logs decision trajectories (*suggest* → *accept/decline* → *rationale*) to support reflection and prevent structural rigidity, harnessing AI benefits while preserving creative flexibility.

3.5.4 DG4. Voice-authentic delegation with anti-formalization guardrails. To support granular delegation (F5) while preventing the official language barrier identified in F2, enable precise "Send to AI" actions for specific micro-tasks including transitions, connective tissue, diction refinement, and consistency checking, while actively preventing the "official language" problem. All AI contributions return as explicit, reversible insertions that writers can easily identify and modify. The system implements declarative guardrails that encode individual voice traits, informal speech patterns, and emotional authenticity markers. When AI generates overly formal, institutionalized language that contradicts the writer's established voice, the system flags these issues and offers alternative approaches that preserve human conversational authenticity. This design enables efficient AI assistance while rigorously safeguarding against the systematic voice erosion that occurs when AI imposes formal linguistic structures on personal creative expression.

3.5.5 DG5. Project-persistent memory with pacing diagnostics. Addressing the long-context continuity challenges (F5), establish robust **project-level memory** that maintains continuity of entities, narrative arcs, creative decisions, and adopted constraints across all writing sessions. This memory system prevents the context resets that frustrate writers when working with current AI tools. Additionally, provide **rhythm and duplication diagnostics** including scene-level pacing heatmaps and repetition detection flags. These diagnostic tools surface potential issues to both Virtual Readers and writers themselves, enabling proactive narrative tightening or strategic expansion before problems compound. This design addresses the long-context fragility and repetitive content generation that limit effective AI collaboration.

4 Plotania System Design

Plotania integrates three core components: a contextual editor with embedded AI features, Virtual Readers providing genre-calibrated feedback, and transparent attribution (Figure 2).

4.1 Interface Architecture

Plotania implements an editor-based architecture addressing writers' need for "light scaffolding" (F1) while avoiding rigid structures

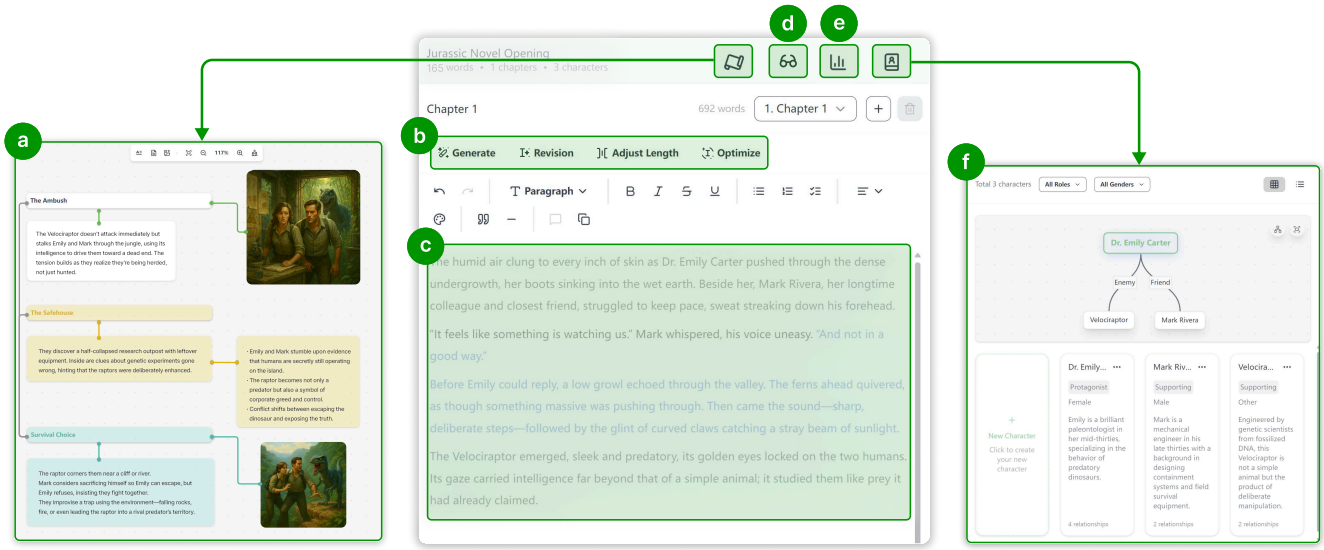

Figure 2: Plotania System Overview. Complete writing environment integrating six core components: (a) story canvas, (b) AI functions, (c) main editor, (d) virtual reader, (e) real-time attribution statistics, and (f) character dashboard.

that constrain exploratory processes. The contextual editor enables AI collaboration without disrupting natural writing workflows.

Editor with Embedded AI Features: The primary interface embeds AI interactions directly within document context, eliminating repeated context reestablishment. Writers invoke AI assistance at any text location through contextual menus, receiving suggestions maintaining full awareness of narrative elements, character development, and story constraints. Four specialized AI functions integrate seamlessly (Figure 3):

- **Generate**: Creates new content and continues existing narratives as negotiable drafts rather than fixed templates (DG3), maintaining continuity with established story elements.
- **Revision**: Provides comprehensive analysis across grammar, style, clarity, and coherence while maintaining authorial voice and intent.
- **Adjust Length**: Handles precise expansion/condensation tasks including transitions and connective tissue that writers find "necessary but creatively unfulfilling" (F5).
- **Optimize**: Enhances text with anti-formalization guardrails preventing the "official language problem" while preserving authentic voice (DG4).

Plot Canvas: Supporting writers' preference for "plain text so I can change it anytime" (F1), provides flexible planning where AI-generated outlines remain negotiable drafts. Writers can expand, contract, reorder, and replace elements without constraint.

Character Dashboard: Maintains persistent character profiles, relationships, and arcs across sessions, addressing "project-level memory" needs (F5) for narrative coherence in longer works.

Project-Aware AI Collaboration: Unlike traditional chat interfaces losing context between sessions, Plotania maintains full project awareness—characters, plots, themes, and constraints—across all interactions, preventing "repetitive content or narrative degradation" (F5).

4.2 User Scenario Walkthrough

Figure 4 illustrates Plotania's collaborative workflow through a complete writing session.

The scenario illustrates three key aspects of Plotania's collaborative workflow: **Progressive User Control**, where writers initially leverage AI assistance for ideation (30% user) but gradually take greater creative control through iterative refinement (65% → 90% user); **Contextual Feedback Integration**, where Virtual Reader provides genre-calibrated comments at natural breakpoints without interrupting creative flow; and **Transparent Attribution Tracking**, where real-time attribution statistics reflect the dynamic balance between AI scaffolding and human creativity. Each interaction maintains full project awareness, ensuring that AI suggestions remain story-consistent, Virtual Reader comments reference previous developments, and attribution tracking preserves complete editing history.

4.3 Virtual Reader: Genre-Calibrated Audience Feedback

While the editor interface enables efficient AI collaboration, our formative study revealed that writers need sophisticated audience feedback beyond generic AI suggestions. The Virtual Reader addresses genre-specific audience calibration failures (F3) by implementing DG2: replacing generic AI suggestions with targeted feedback from specialized reader perspectives that understand genre conventions, emotional tones, and reader community expectations (Figure 5).

Figure 3: AI Function Workflow. Users submit writing requests, select appropriate AI functions, and review generated results. Four core functions: (a) Generate new content, (b) Revision for improvement, (c) Adjust Length for expansion/condensation, and (d) Optimize for enhancement.

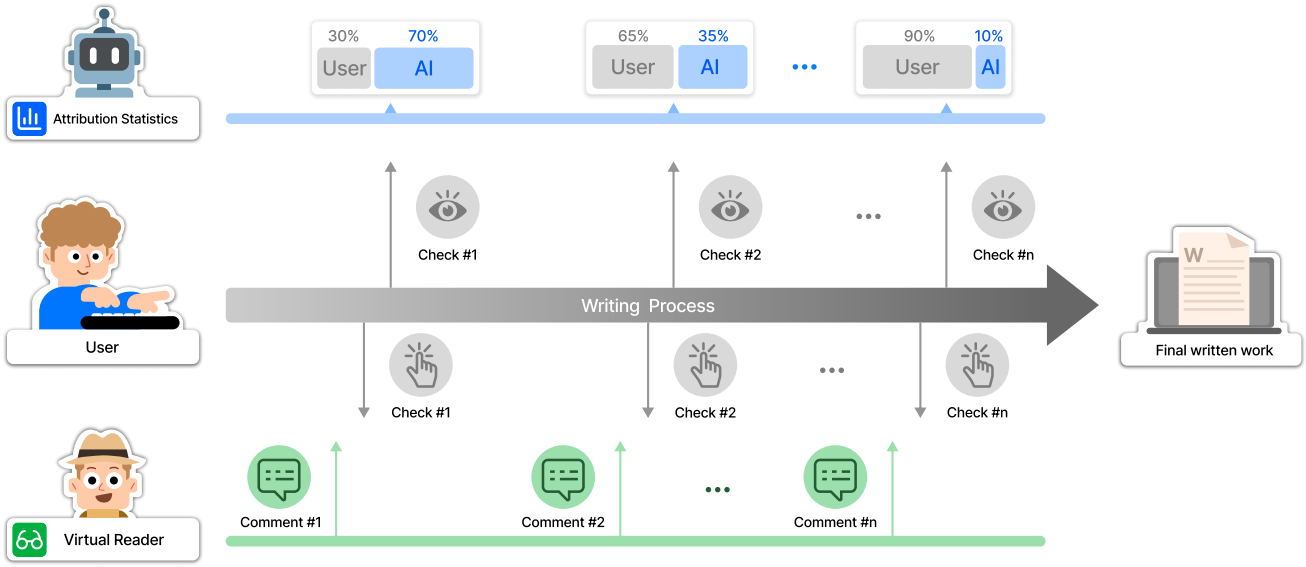

Figure 4: User Scenario Walkthrough. Iterative collaborative writing process showing dynamic attribution evolution and integrated feedback loops.

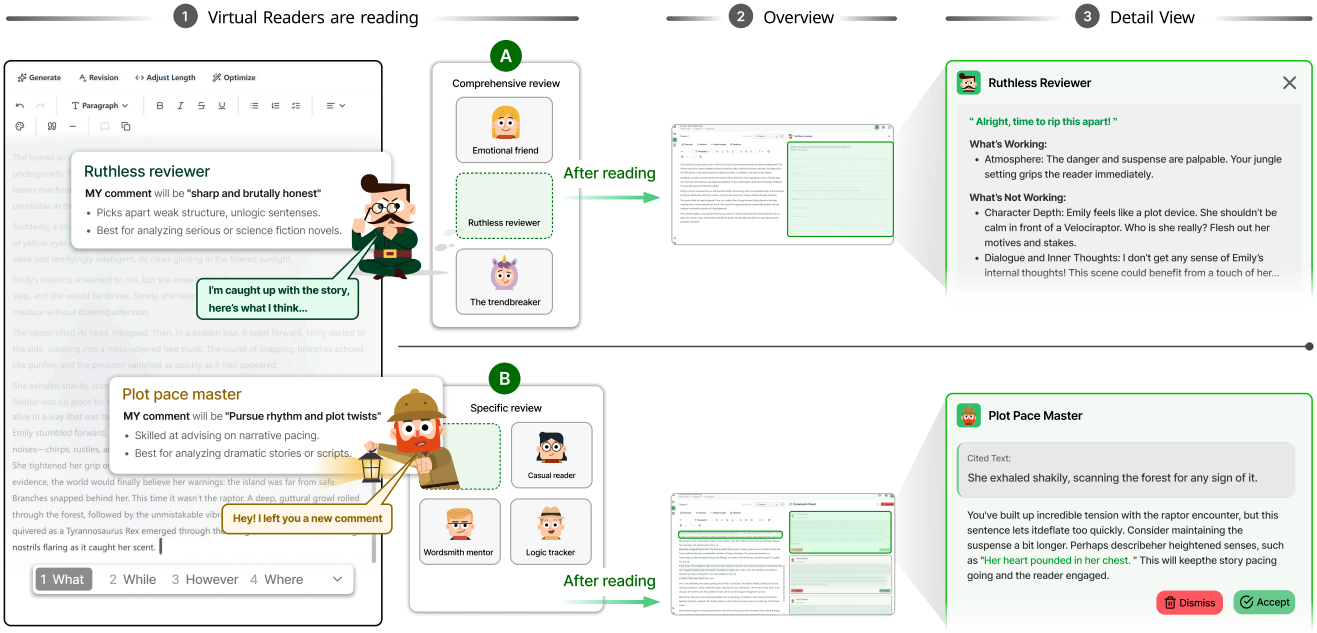

Figure 5: Virtual Reader Workflow. Proactive three-stage reading process: (1) active reader engagement with ongoing text analysis, (2) comprehensive overview with multiple reader perspectives offering (A) comprehensive review and (B) specific review modes, and (3) detailed feedback delivery with specific suggestions and user controls.

Reader-Based Architecture. Virtual Reader implements specialized readers with distinct expertise areas, evaluation priorities, and feedback styles (Table 1), transforming AI feedback from generic writing advice to specialized reader perspective simulation.

Controllable Feedback Architecture. Virtual Reader provides control mechanisms addressing writers' need to avoid "emotional disruption of unfiltered reader comments" while accessing "incisive, role-specific feedback":

- **Content-Based Triggers**: Automatically engages at content thresholds, providing feedback at natural narrative breakpoints
- **Time-Based Intervention**: Offers contextual suggestions after writing intervals while respecting creative flow
- **User State Awareness**: Monitors writing behavior to determine appropriate intervention timing and intensity
- **Granular Scope Control**: Writers configure feedback depth from light structural suggestions to comprehensive analysis
- **Evidence-Anchored Reasoning**: Every comment cites specific textual evidence with genre-appropriate reasoning

Context-Aware Feedback Delivery. Virtual Reader integrates with Plotania's context-aware interaction model:

- **Contextual Document Annotations**: Direct highlighting and comments at specific locations
- **Integrated Feedback Coherence**: Maintains consistency between high-level story development and textual implementation
- **Reader-Specific Intervention Patterns**: Each reader type employs distinct intervention strategies

Implementation. Virtual Readers are implemented using OpenAI's GPT-4o (temperature=0.7, max_tokens=3000) with carefully engineered prompts designed through iterative refinement with fiction writers. Each persona employs distinct voice characteristics, evaluation priorities, and feedback delivery styles to simulate authentic reader perspectives. Comprehensive mode generates holistic narrative feedback in natural language, while specific mode produces structured JSON output mapping text segments to targeted comments with cited evidence. Complete persona prompts and technical specifications are provided in Appendix A.

4.4 Transparent Attribution

Beyond providing sophisticated AI assistance and feedback, maintaining writers' sense of authorship is critical for creative agency. To address authorship anxiety and identity dilution concerns (F2), Plotania implements DG1 through a comprehensive attribution system providing complete visibility while preserving writers' creative ownership.

Impact-Based Attribution Framework. The system evaluates attribution by **creative significance** rather than word counts, through a six-dimensional taxonomy (Table 2).

Implementation. The attribution algorithm employs GPT-4o-mini (temperature=0.1) to analyze text differences given operation context (expand, optimize, compress), classifying segments by change type, authorship, and position. Operation-based weighting

Table 1: Virtual Reader Categories

Category	Reader	Specialized Focus & Community
Structural	Ruthless Reviewer	Publication standards, plot development, pacing
	Logic Tracker	Plot coherence, consistency, world-building
	Plot Pace Master	Story rhythm, tension flow, chapter hooks
Emotional	Emotional Friend	Character connection, relationship dynamics
	Casual Reader	Natural reading experience, relatability
Creative	Trendbreaker	Originality, cliché detection, creative risks
	Wordsmith Mentor	Language craft, prose quality, voice

Table 2: Attribution Taxonomy: Six-Dimensional Creative Contribution Framework

Priority	Dimension	Creative Contributions
High Weight	Creative Ideation	Original plot concepts, character creation, thematic directions, world-building elements
	Content Development	Expanding outlines into scenes, dialogue creation, descriptive passages, narrative implementation
	Logic Repair	Resolving plot holes, character inconsistencies, setting contradictions, timeline conflicts
Standard Weight	Style Crafting	Writing techniques, narrative voice, prose rhythm, genre conventions
	Quality Enhancement	Grammar corrections, word choice refinement, sentence clarity, detail enrichment
	Readability Enhancement	Pacing adjustments, emotional flow, comprehension improvements, engagement optimization

reflects creative impact: expansion assigns high AI weight (0.9) to added content; optimization assigns moderate weight (0.8) to modifications; compression assigns high weight (0.9) to deletions. This addresses the limitation that simple similarity metrics cannot distinguish substantive contributions from surface refinements. Detailed pseudocode and worked examples are provided in Appendix B.

Real-Time Transparency Architecture. The system implements sophisticated tracking at multiple granularity levels:

Operation-Level Tracking:

- Every modification recorded with timestamp, authorship, and change type
- LLM-based analysis assesses narrative impact using story-level context
- Importance weighting based on influence on plot, character, theme, and style
- Session tracking organizes modifications by editing sessions

Visualization System:

- **Color-coded highlighting**: Visual distinction between human and AI attributions
- **Proportion dashboard**: Quantitative charts showing attribution distribution

Reversible Edit Lineage. Writers can trace and undo any modification at any granularity level, ensuring AI attributions remain explicitly visible and controllable.

Final Decision Attribution. Regardless of collaborative iterations, the final decision maker receives attribution credit, ensuring clear authorship assignment when writers extensively modify AI-generated content.

5 User Study

To address RQ3 and evaluate how our proposed transparency mechanisms, which include attribution visualization and virtual reader feedback, affect creative agency and authorial ownership in AI-assisted writing, we conducted a controlled laboratory study using a within-subjects design.

5.1 Experimental Design

We employed a 2×2 within-subjects factorial design with two independent variables: *Attribution Visualization* (present vs. absent) and *Virtual Reader Feedback* (present vs. absent). The within-subjects design was chosen to control for individual differences in writing ability and AI familiarity while maximizing statistical power. This resulted in four experimental conditions:

- **C1 (Baseline)**: No attribution visualization, no virtual reader feedback
- **C2 (Attribution Only)**: Attribution visualization enabled, no virtual reader feedback
- **C3 (Reader Only)**: No attribution visualization, virtual reader feedback enabled
- **C4 (Full Features)**: Both attribution visualization and virtual reader feedback enabled

To control for order effects and ensure ecological validity of story development, we used a 4×4 Latin square counterbalancing design. Participants were randomly assigned to four groups (S1-S4, n=5 each), with each group experiencing conditions in a different predetermined order. All participants progressed through the same four narrative phases in sequence (opening → development → climax → conclusion), but the experimental condition varied systematically

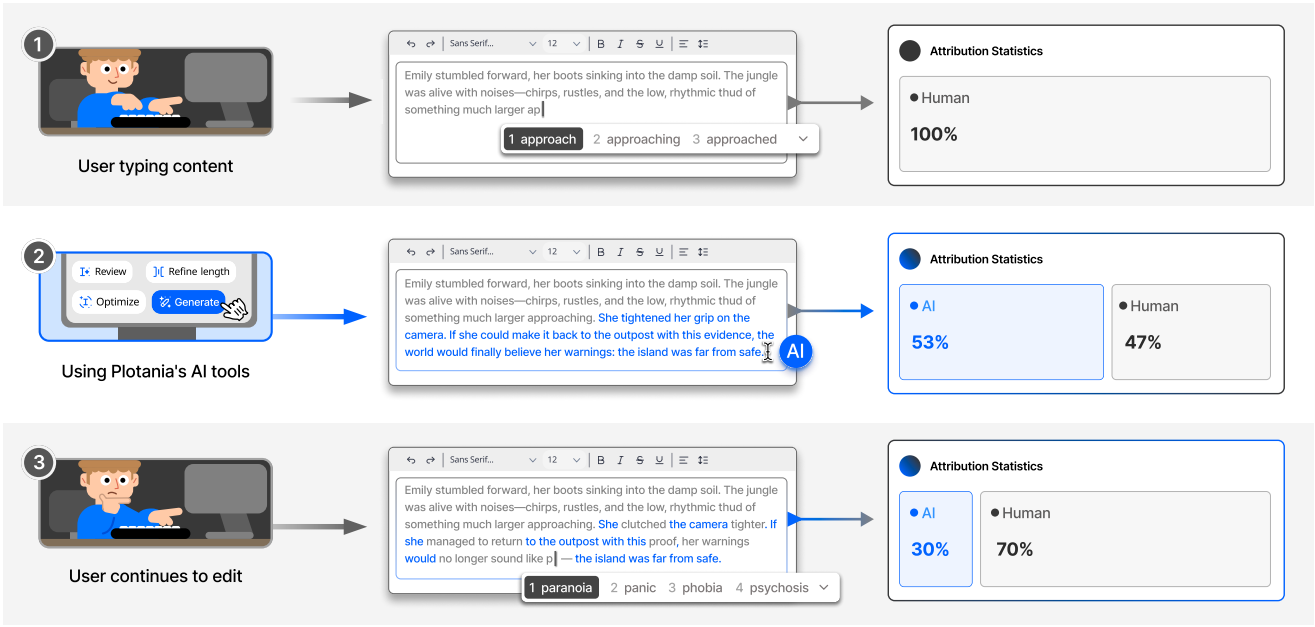

Figure 6: Attribution Tracking Workflow. Real-time demonstration of dynamic attribution changes during collaborative writing.

across groups. This design ensured that each condition appeared equally often in each serial position *and* each narrative phase, controlling for potential learning effects, fatigue, and phase-specific variance while enabling analysis of how transparency effects vary across different stages of story development.

5.2 Participants

We recruited 20 participants (12 female, 8 male, ages 22–32, M=26.4, SD=3.2) through academic institution mailing lists and social media. Writing experience ranged from less than 1 year to over 10 years: 2 participants had less than 1 year of experience, 9 had 1-3 years, 1 had 3-5 years, and 8 had 5-10 years (M=3.8 years). The vast majority (18 out of 20) had more than 1 year of creative writing experience, enabling them—with AI assistance—to produce high-quality creative work within the limited 12-minute time frame. Among participants, 14 primarily wrote short fiction or stories, while 6 regularly wrote essays or diary entries, representing a mix of hobbyist and semi-professional writers.

Our target population comprised emerging and hobbyist writers who actively use AI tools for creative writing—a deliberate choice that differs from the formative study's more experienced fiction writers (6/10 with 4+ years of serialization experience). This sampling strategy reflects our design goal of supporting writers who are still developing their craft and may benefit most from transparency mechanisms to navigate AI collaboration while building authorial confidence. All participants (100%) reported prior experience with AI writing tools: 4 used them daily, 8 frequently, and 8 occasionally. The most commonly used tools were ChatGPT, Gemini, and Claude. Participants were compensated $25 for approximately 90 minutes of participation.

5.3 Task and Procedure

The creative writing task was designed to balance ecological validity with experimental control. Each participant completed four 12-minute writing sessions using Plotania, with sessions structured to follow natural story progression: opening (establishing characters and setting), development (building conflict), climax (peak tension), and conclusion (resolution). The 12-minute duration was chosen based on pilot testing to allow sufficient creative output while preventing fatigue effects.

Participants selected one of four predefined themes (Campus Mystery, Workplace Challenge, Emotional Conflict, Sci-Fi Adventure) designed to offer similar creative affordances while appealing to diverse interests. Predefined themes enhanced experimental control, allowing observed differences to be attributed to transparency features rather than topic familiarity. Theme distribution was balanced: 5 Campus Mystery, 7 Workplace Challenge, 6 Emotional Conflict, 2 Sci-Fi Adventure.

The experimental session lasted approximately 90 minutes per participant and followed this protocol:

(1) **Pre-study**: Informed consent, demographic questionnaire, and system introduction (10 min)
(2) **Writing Sessions**: Four rounds of 12-minute writing tasks with different conditions (48 min)
(3) **Post-task Questionnaires**: Brief experience survey after each writing session (12 min)
(4) **Post-study Interview**: Semi-structured interview about overall experience (20 min)

During writing sessions, we recorded system interactions, writing behaviors, and usage patterns for AI assistance features. Between sessions, participants completed questionnaires measuring perceived creativity, sense of ownership, writing satisfaction, and system usability (see Appendix C for complete questionnaire items).

5.4 Measures

We collected both quantitative and qualitative data:

Behavioral Metrics:

- *Writing productivity*: Total word count and writing rate per session
- *AI assistance patterns*: Frequency and type of AI feature usage (generate, revision, adjust length, optimize)
- *Attribution engagement*: Time spent viewing attribution highlights, number of attribution checks per session
- *Reader feedback interaction*: Response rates to reader suggestions (accept, dismiss, modify), time spent on reader comments

Subjective Measures: All items measured on 7-point Likert scales (1=strongly disagree, 7=strongly agree):

- *Creative Agency*: 3-item scale measuring perceived control over creative decisions and story direction
- *Authorial Ownership*: 3-item scale measuring sense of ownership over the final work and ability to distinguish personal vs. AI attributions
- *Audience Awareness*: 3-item scale measuring consideration of reader perspectives during writing
- *Cognitive Load*: 3-item scale measuring mental effort and interruption to creative flow (reverse-coded items included)

Qualitative Data: Semi-structured interviews explored participants' experiences with transparency features, including their impact on creative confidence, AI collaboration strategies, and long-term usage intentions.

5.5 Data Analysis

For quantitative data, we employed repeated measures ANOVAs with Attribution Visualization (2 levels) and Virtual Reader Feedback (2 levels) as within-subjects factors. Effect sizes are reported as partial eta-squared (η_p^2) and Cohen's d, with Cohen's conventions for interpretation (small: 0.01, medium: 0.06, large: 0.14 for η_p^2; small: 0.2, medium: 0.5, large: 0.8 for d). Post-hoc pairwise comparisons used Bonferroni correction to control family-wise error rate (α = .017). For behavioral metrics, we additionally analyzed temporal patterns using time-series analysis where appropriate.

Statistical Power. Post-hoc power analysis revealed that our within-subjects design (N=20, 4 conditions, Bonferroni-adjusted α = .017) provided approximately 80% power to detect medium-to-large effects (Cohen's f $\geq$ 0.44, equivalent to d $\geq$ 0.88). However, for the small effect sizes observed in creative agency measures (d = 0.13), our achieved power was approximately 3%. We thus frame our quantitative findings as **exploratory**, emphasizing effect size patterns and directional trends rather than definitive causal claims, with qualitative findings providing crucial interpretive depth.

Qualitative interview data underwent systematic thematic analysis following Braun and Clarke's six-phase approach. Two researchers independently coded all transcripts, achieving satisfactory inter-rater reliability (Cohen's $\kappa > 0.80$). Themes were developed iteratively through constant comparison and validated through member checking with a subset of participants. All sessions were video-recorded and transcribed verbatim for detailed behavioral coding.

6 Results

Our within-subjects experiment with 20 participants examined how attribution visualization and virtual reader feedback affect creative agency in AI-assisted writing. This section presents both quantitative behavioral measures and qualitative interview findings exploring how participants maintained authorial ownership and audience awareness while navigating the tension between AI assistance and creative control.

6.1 Quantitative Results

6.1.1 Writing Productivity and AI Usage Patterns. **Writing Productivity and AI Usage.** Attribution-only (C2) generated highest median word count (Mdn=1,699, +143 words over Baseline Mdn=1,556), while Full System (C4) produced fewer words (Mdn=1,472). Transparency mechanisms reduced AI usage across all conditions compared to baseline (C1: M=3.0 uses), with Attribution-only showing largest reduction (-0.6 uses), suggesting transparency encourages *self-reliant writing behavior*.

6.1.2 Subjective Experience Measures. **Creative Agency Effects.** Creative agency (α=.77) showed no statistically significant differences after Bonferroni correction ($\chi^2(3)$=1.551, p=.671). Full System achieved highest scores (M=5.44), followed by Baseline (M=5.31), Attribution-only (M=5.26), and Reader-only (M=5.19). Effect size patterns reveal design implications: individual features produced mixed results (Attribution: d=-0.05, Reader: d=-0.13), but combination generated modest enhancement (d=+0.13), suggesting transparency mechanisms create *complementary* rather than additive effects.

Audience Awareness. Reader-only (C3) produced highest awareness scores (M=4.75, α=.65), with 0.42-point increase from baseline, though not statistically significant ($\chi^2(3)$=1.596, p=.660). This gain comes with a 0.12-point agency reduction, revealing fundamental tension between social awareness and authorial autonomy.

Cognitive Load and Flow. Cognitive load remained stable across conditions (range: M=5.23-5.52). However, virtual reader feedback reduced creative flow (C3: M=5.05 vs C1: M=5.85), indicating transparency's cost lies in creative rhythm disruption, not mental effort.

Attribution Clarity. Attribution features showed strong effects on distinguishing human vs AI contributions (d=0.87). Attribution-only achieved highest clarity (M=5.95) vs Baseline (M=5.00).

Preference-Performance-Satisfaction Dissociation. Post-study preferences revealed a striking three-way dissociation (Figure 7): while 80% preferred Full System overall, 55% found Reader-only most creatively satisfying despite its lowest agency scores.

Feature-level evaluations reinforce this paradox: virtual reader feedback received higher ratings (M = 5.55/7) than attribution visualization (M = 4.60/7), yet attribution conditions showed higher agency scores. This indicates *utility* and *enjoyment* operate as distinct constructs in creative AI evaluation.

Table 3 provides comprehensive descriptive statistics. Key effect sizes: audience awareness (+0.38 with Reader), attribution clarity (+0.87), creative agency (+0.13 with Full System). These patterns should be interpreted cautiously given limited statistical power.

Phase-Based Patterns. Mixed-effects models revealed a notable Reader × Conclusion interaction for audience awareness (β=+2.20, p=.038, not significant after Bonferroni correction). Creative agency trajectories (Figure 9A) showed Reader-only peaked during Development (M=5.80) but declined at Conclusion (M=4.70), while Baseline achieved highest scores at narrative culmination (Conclusion: M=6.00). These exploratory patterns suggest transparency may exhibit phase-dependent effects.

6.2 Qualitative Findings

Post-session interviews revealed four major themes regarding participants' experiences with transparency mechanisms.

6.2.1 Virtual Reader Feedback as Multi-Dimensional Creative Support. Our analysis reveals that virtual readers function as **contextual creative partners** rather than simple feedback providers, creating value across three interdependent dimensions:

Cognitive Support Through Professional Expertise: Virtual readers provided technically sophisticated feedback valued for contextual awareness. Critically, we observed **sustained behavioral learning**: after Plot Pace Master alerted P12 about "overly dense" plot information, P12 subsequently paused four times during writing to self-check narrative pacing. P1 reported: "Ruthless Reviewer's suggestions will continue influencing how I construct narrative structure." These patterns suggest virtual readers function as **internalized writing coaches** fostering individualized skill development.

Social Support Through Reader Differentiation: The distinct reader types enabled different creative dialogues. P17 appreciated: "The plot pace master's suggestions in dialogue format made it feel like talking to a person." P20 found complementary roles: "The casual reader acts like a supporter... the literary craftsperson helps refine expression."

Emotional Support Through Creative Companionship: Six participants spontaneously expressed strong affection for specific Virtual Readers. P16: "I absolutely love the Casual Reader character... reducing much writing pressure and giving me more confidence." P12 valued "companionship, because creating alone can be quite lonely." P9 emphasized the "emotional value" from virtual readers, noting "it feels like having someone who cares about my story." These findings reveal **parasocial creative companionship** as a distinct value dimension.

6.2.2 Attribution Transparency: From Identity Disruption to Editorial Control. Our analysis reveals transparent attribution as a psychologically transformative experience that challenges, then reshapes, writers' relationship with AI assistance through three distinct phases.

Identity Disruption and Anxiety. Attribution visualization initially created intense **creative identity tensions** that challenged writers' fundamental self-concept. P15 expressed deep frustration: "I wrote about 300-400 words in the outline... but it shows AI contributed more than I did. This made me very dissatisfied." This computational mismatch between user perception and system calculation created what P11 described as writing anxiety: "When AI-generated parts exceed what I wrote myself, I actually feel a bit anxious... wondering if this article was truly written independently by me." P14 extended this anxiety to social validation, worrying: "If the algorithm thinks AI-generated parts exceed what I wrote myself, would reviewers on publishing platforms also think I wrote purely AI-generated content?"

Gradual Adaptation Process. Rather than immediate rejection, participants developed sophisticated coping strategies that transformed their relationship with transparency. P11 articulated a temporal adaptation: "I would use it moderately... In the early stages when inspiration is limited, seeing mostly AI-generated content creates anxiety and disappointment. But I would use it more in later stages when my own thoughts increase." This selective engagement reflects what P1 described as motivational reframing: "It affects me - if I see AI contributing a lot, I want to write more myself," suggesting transparent attribution can motivate rather than discourage creative effort.

Reframing as Editorial Empowerment. Ultimately, participants discovered attribution's unique value as a **revision and quality control tool**. P6 developed a sophisticated analogy: "Like a plagiarism detection system for papers, it lets me know which novel ideas and thoughts are represented in the black (Human) markings, showing my creative flow is in good condition." P18 contrasted this capability with mainstream AI tools: "In popular models we could not differentiate like a which text is written by the AI," noting this critical gap in ChatGPT, Claude, and similar tools. Most significantly, participants reframed attribution as an **editing affordance**, with P18 describing using it to "differentiate easily and remove the plagiarism in your writing." This instrumental use for selective revision and quality control ultimately enabled rather than constrained creative agency, transforming anxiety-inducing visibility into empowering editorial control.

7 Discussion

Our study reveals a fundamental paradox: despite high engagement with virtual reader feedback and accurate attribution information, transparency mechanisms demonstrated *complex and counterintuitive effects* on perceived creative agency (**RQ3**), challenging assumptions that transparency automatically improves human-AI collaboration. Participants experienced tensions as virtual readers transformed individual authorship into collaborative performance while attribution visualization created identity anxiety (**RQ1**), leading writers to develop protective strategies including selective AI engagement and cognitive reframing (**RQ2**). These findings suggest transparency in creative contexts requires fundamentally different design principles that preserve creative identity alongside providing information.

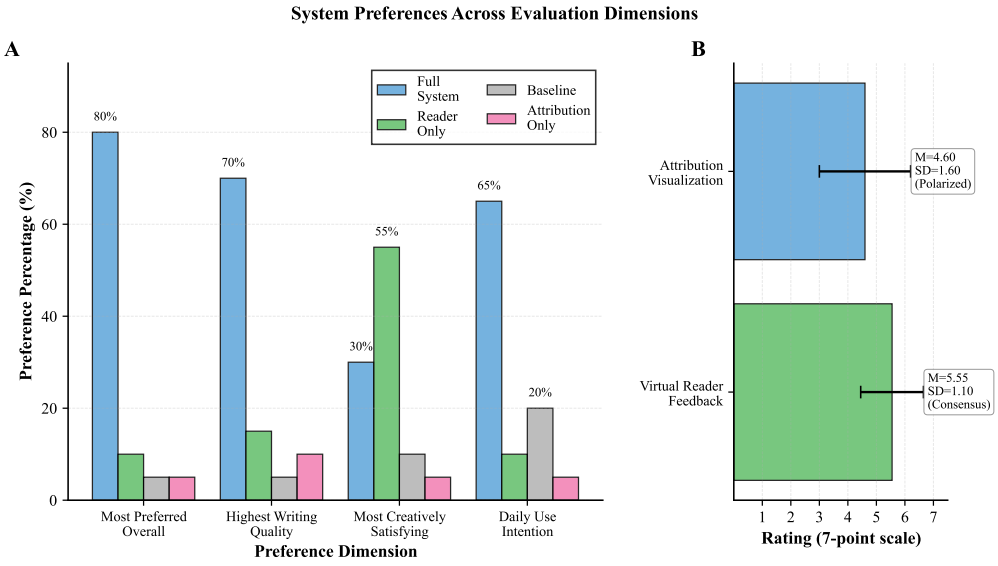

Figure 7: System preferences across evaluation dimensions. Full System dominates in overall preference (80%) and writing quality (70%), while Reader-only leads in creative satisfaction (55%). Virtual Reader Feedback receives higher consensus ratings (M=5.55) compared to Attribution Visualization's polarized responses (M=4.60).

Table 3: Subjective Experience Measures by Condition

Measure	C1 Baseline	C2 Attribution	C3 Reader	C4 Full	Range	Effect Size
Primary Composite Measures						
Creative Agency	5.31±0.99	5.26±0.91	5.19±0.88	5.44±0.91	0.25	**+0.13**
Audience Awareness	4.33±1.09	4.60±1.15	4.75±1.13	4.60±1.05	0.42	**+0.38**
Cognitive Load	5.37±1.05	5.52±1.02	5.23±1.39	5.28±1.04	0.28	+0.14
Key Individual Components						
Story Control	5.40±1.14	5.50±1.05	5.25±0.91	5.85±0.75	0.60	+0.45
Decision Driven	5.45±1.15	5.55±1.15	5.65±0.81	5.95±1.00	0.50	+0.50
Reader Perspective	4.10±1.25	4.45±1.39	4.70±1.53	4.45±1.28	0.60	+0.60
Reader Understanding	4.55±1.23	4.75±1.25	4.80±1.20	4.75±1.12	0.25	+0.25
Process and Control Measures						
Attribution Clarity	5.00±1.17	5.95±1.00	5.55±1.15	5.85±1.23	0.95	**+0.95**
Flow Interruption†	5.85±1.23	5.65±0.93	5.05±1.54	5.05±1.43	0.80	**-0.80**
Feedback Clarity	4.75±1.68	5.20±1.32	5.20±1.36	5.20±0.95	0.45	+0.45

Notes: All measures use 7-point Likert scales (1=strongly disagree, 7=strongly agree). †Flow Interruption: Higher scores indicate better flow (reverse-coded item). **Bold effects** indicate largest positive/negative changes from baseline. Effect Size column shows largest change from baseline across conditions.

7.1 Virtual Readers as Creative Companions

Virtual reader feedback created a fundamental tension: while participants engaged extensively with the system, their perceived creative agency showed a modest but consistent decrease, revealing that *utility* and *agency* are distinct outcomes in creative AI systems. This reduction stems from two mechanisms. First, proactive feedback disrupted creative flow, forcing participants to alternate between generative and evaluative modes rather than maintaining the immersive state essential for creative work [57]. Second, different

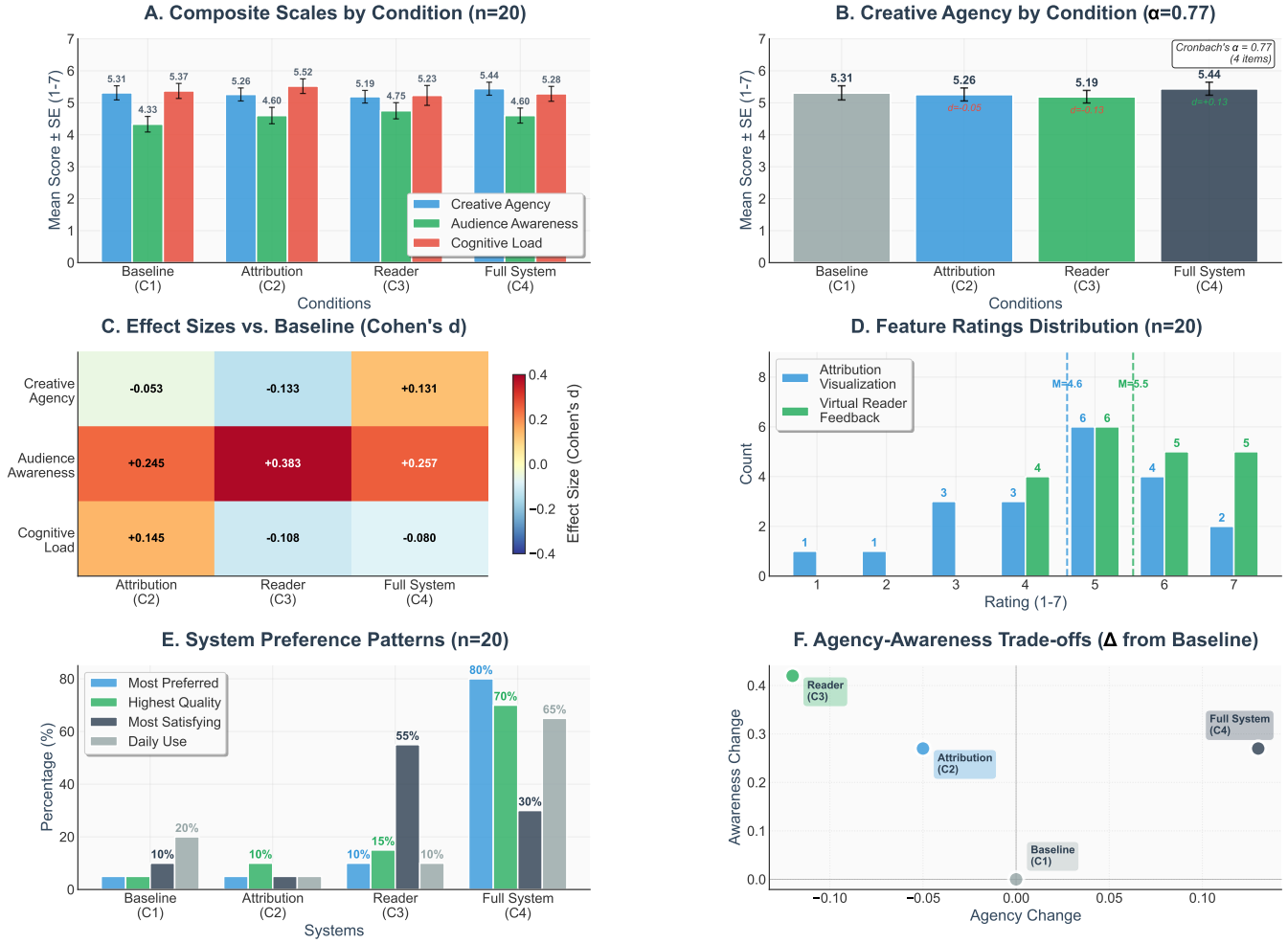

Figure 8: Statistical analysis of transparency effects across six panels: (A) Creative agency, audience awareness, and cognitive load composite scores across conditions, (B) Creative agency detailed view showing Full System enhancement (d=+0.13), (C) Effect size heatmap for composite scales showing moderate positive effects for audience awareness and modest negative effects for agency, (D) Feature rating distributions, (E) Preference patterns revealing 80% prefer Full System overall but 55% find Reader-only most satisfying, (F) Trade-off analysis quantifying agency-awareness exchange patterns.

Virtual Readers created asymmetric psychological impacts: evaluative readers like Ruthless Reviewer generated stronger agency disruption through critical judgment, while supportive readers like Emotional Friend preserved more creative autonomy despite similar interaction frequency. By making the imagined audience persistently present, virtual readers transformed the author-audience relationship from an internal dialogue into externalized performance evaluation, fundamentally reframing individual authorship as collaborative performance management.

Most surprisingly, virtual readers created *parasocial relationships* [43] that provided emotional companionship, with participants seeking "companionship" (P12) and "emotional value" (P9) that sustained motivation. P4's desire for "barrage comments" and feeling "relaxed like someone's presence" reveals how AI personas function as *creative companions* through what we term *dynamic parasocial engagement*—relationships that felt bidirectional and responsive to specific creative choices [51, 60]. The preference for dialogue-style feedback suggests conversational AI interfaces may be more psychologically sustainable by satisfying social connection needs during solitary creative work.

Critically, our observations suggest virtual reader feedback fostered **individualized craft development** rather than stylistic homogenization. Writers internalized feedback in domain-specific ways: P12 developed sustained pacing awareness through self-initiated rhythm checks; P1 restructured approaches to character motivation. The diversity of Virtual Readers preserved stylistic plurality by enabling selective engagement matching individual developmental needs, suggesting virtual readers promote **personalized skill acquisition** within a supportive framework respecting authorial autonomy.

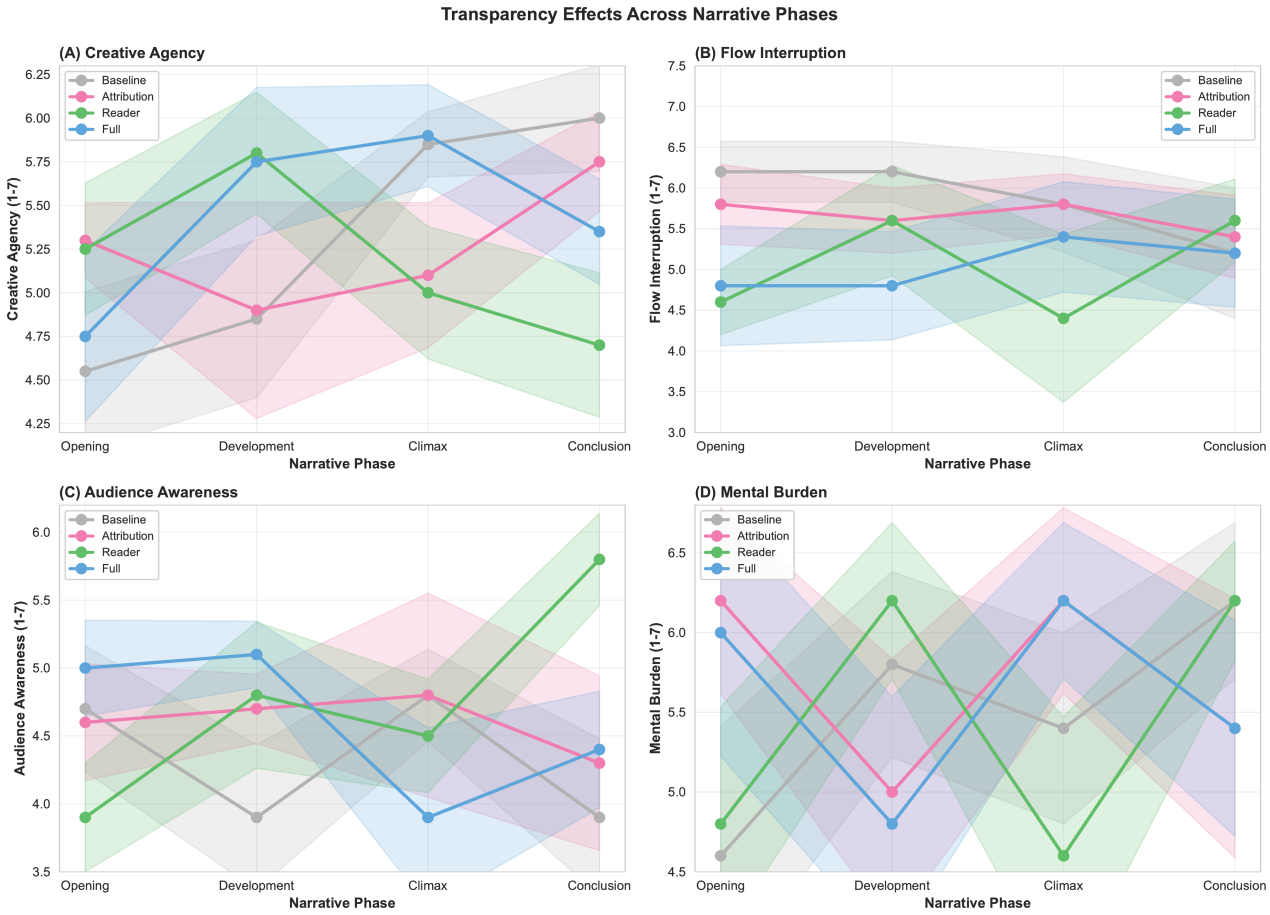

Figure 9: Transparency effects across narrative phases reveal phase-dependent patterns. Reader-only creative agency peaks at Development but declines at Conclusion; Baseline achieves highest scores at Climax/Conclusion. Notable Reader × Conclusion interaction for Audience Awareness (p=.038, uncorrected). Error bands=SEM; N=5 per Condition × Phase cell.

However, despite modest agency reduction, participants demonstrated **sustained adoption intention**, revealing preference-performance dissociation. These findings challenge assumptions that visibility enhances human agency [28], revealing **agency sensitivity** where even modest transparency interventions affect creative control, suggesting creative AI transparency must be *agency-preserving*, not merely *informative*.

7.2 Transparency Paradoxes and Design Challenges

Attribution visualization created *attribution anxiety* despite technical success. P15's frustration exemplifies this: "I wrote about 300-400 words in the outline... but it shows AI contributed more than I did." Our system measured textual similarity rather than creative agency, rendering invisible what we term ***invisible creative labor***—ideational direction, conceptual framing, and aesthetic judgment. P10's insight reveals the inadequacy of binary human-AI attribution: "AI writing is also done under my operation, so it can't be seen as completely written by AI."

This created *avoidance motivation*: users modified behavior to achieve favorable scores rather than optimal creative outcomes. Participants also expressed *anticipatory algorithmic judgment*—fears about how external systems would evaluate their AI usage. P14's concern about "publishing platforms" reveals how attribution systems become entangled with broader creative legitimacy networks, potentially creating a *transparency trap* where writers minimize AI usage to avoid algorithmic discrimination rather than preserve creative integrity [19].

The "algorithmic anxiety" [15] reveals we should ask "What happens to human creative identity when AI contributions become visible?" rather than simply making contributions visible.

Importantly, participants exhibited **gradual adaptation** over time. Initial anxiety gave way to more nuanced engagement as writers learned to interpret attribution information as editorial guidance rather than judgment. P6 reframed attribution as "like a plagiarism detection system, it lets me know which novel ideas are mine," transforming anxiety into editorial control.

However, our findings reveal **complementary effects**: combined transparency features mitigate rather than amplify individual

limitations. The Full System balanced emotional satisfaction of virtual readers with control benefits of attribution, creating a framework that preserves creative agency while providing transparency benefits. Current attribution approaches represent a category error in that they apply measurement frameworks developed for task-oriented AI to domains where human identity and meaning-making are central.

7.3 Multi-Dimensional Creative AI Experience

Our findings reveal a striking **preference-performance dissociation**: while 80% preferred the Full System overall, 55% found Reader-only most creatively satisfying despite its lowest agency scores. This three-way dissociation—between rational evaluation (what works), emotional satisfaction (what feels good), and behavioral intention (what they would use)—indicates that *utility* and *enjoyment* operate as distinct constructs in creative AI evaluation. Features that feel helpful may not necessarily feel empowering, suggesting traditional satisfaction metrics inadequately capture user experience in identity-central domains.

Phase-Dependent Creative Autonomy. Our phase analysis reveals transparency mechanisms that scaffold exploration become autonomy threats during resolution. Virtual reader feedback achieved highest agency during development (M=5.80) but plummeted at conclusion (M=4.70)—a 1.10-point decline representing the sharpest phase-specific drop observed. This *Narrative Autonomy Hypothesis* suggests writers' tolerance for external intervention varies with narrative stakes—exploration phases welcome feedback, but conclusion phases demand *crystallization of authorial intent* where external input challenges the writer's right to determine meaning. As baseline conditions achieved their highest agency scores precisely during these critical junctures (Climax M=5.85, Conclusion M=6.00), writers appear to recognize and protect their autonomy needs at story culmination points. Effective systems must implement *phase-adaptive transparency*.

7.4 Adaptive Strategies for AI Collaboration

Despite transparency-induced agency challenges, participants developed protective strategies to maintain creative agency: **Selective AI Engagement** (strategically limiting AI usage in voice-critical passages while recruiting it for structure [39]), **Reframing AI as Tool** (mentally repositioning AI from "co-author" to "advanced tool"), and **Identity-First Creation** (establishing creative voice before engaging AI assistance). These strategies reveal that writers inherently understand identity threats and develop protective responses without explicit guidance, suggesting successful creative AI systems should support rather than work against these natural behaviors.

7.5 Design Implications: Agency-Preserving Transparency

We propose four principles balancing information provision with creative empowerment:

Layered Attribution: Recognize different contribution types (ideational, directorial, editorial, generative) rather than binary percentages, addressing the *invisible creative labor* problem. A writer who provides the core plot concept, character motivations, and thematic direction has made substantial creative contributions even if AI generates more surface-level text. Attribution systems should weight these higher-order contributions appropriately, perhaps through explicit tagging of contribution types or weighted scoring that privileges conceptual over textual contributions.

Context-Sensitive Transparency: Attribution information should adapt to creative context. During active writing, hide or minimize attribution displays to prevent the measurement anxiety we observed and avoid flow disruption. After natural stopping points such as completing paragraphs, sections, or writing sessions, systems should surface attribution insights when writers shift from generative to reflective modes. Systems should detect writing activity patterns to automatically adjust transparency intensity. This principle recognizes that the same information can be empowering or anxiety-inducing depending on when it appears in the creative workflow.

Collaborative Framing Control: Allow users to adjust "audience presence intensity" from solitary writing to full collaboration. Our phase analysis revealed that writers' tolerance for external input varies dramatically across narrative phases. Early exploration benefits from active audience engagement, while conclusion phases require protected space for authorial crystallization. Systems should provide explicit controls for adjusting feedback frequency, virtual reader visibility, and notification patterns, enabling writers to modulate external presence based on their current creative needs.

Emotionally Intelligent Transparency: Provide emotional support through personified interfaces while maintaining attribution accuracy. The strong parasocial bonds participants formed with virtual readers suggest that transparency mechanisms can be delivered through emotionally supportive channels. Rather than presenting attribution as cold metrics, systems might frame contribution information through encouraging virtual reader commentary, celebrating human creative decisions while maintaining accurate records.

7.6 Limitations

Several critical limitations constrain our findings' generalizability and interpretation:

Statistical Power and Sample Size. Our within-subjects design with N=20 participants was underpowered to detect small effect sizes. Post-hoc power analysis revealed approximately 3% power to detect the observed creative agency difference ($d = 0.13$) between Full System and Baseline conditions, requiring $N \geq 80$ for adequate power (80%) to detect $d = 0.3$ effects. Consequently, our findings should be interpreted as **exploratory** rather than confirmatory. Only attribution clarity effects ($d = 0.87$) achieved adequate statistical power and significance.

Ecological Validity. Laboratory sessions (12 minutes with standardized prompts) differ substantially from authentic creative writing characterized by sustained engagement, personal investment, and iterative revision across days or weeks. Writers may develop different relationships with transparency mechanisms over extended use, and the compressed timeline may have amplified or attenuated certain effects.

Sample Constraints. Our academic participants (N=20, emerging/hobbyist writers with 1-5 years experience) may not represent

professional fiction writers or casual hobbyists. Individual differences in writing expertise, AI comfort, and creative identity may moderate transparency effects in ways our sample size could not detect.

Measurement-Construct Alignment. Our attribution system measured textual similarity rather than the multi-dimensional creative contributions outlined in Table 2, rendering invisible ideational direction, conceptual framing, and aesthetic judgment. This raises a critical question: does the attribution anxiety we observed reflect inherent problems with transparency in creative work, or does it reflect *bad transparency*—metrics misaligned with what writers value about their contributions? This distinction suggests future research must differentiate between transparency mechanisms that accurately reflect creative agency versus those providing misleading visibility.

Cultural Limitations. Our findings assume Western individualistic authorship models emphasizing individual ownership and creative autonomy. Cross-cultural validation is needed in collective creativity cultures where collaborative authorship and shared ownership may be normative.

7.7 Future Work

Our findings establish three transformative research directions:

Creative Companion Paradigm. Moving beyond tool-based AI to relationship-based partnerships that develop persistent creative relationships. Future systems should learn writers' voices, emotional patterns, and creative preferences over extended collaborations. This includes developing AI personas that remember previous projects, understand individual stylistic signatures, and adapt their feedback approaches based on accumulated knowledge of what resonates with each writer. The parasocial relationships observed in our study suggest writers are psychologically prepared for such deeper partnerships.

Adaptive Creative Ecosystems. Systems that understand creative context and identity state, automatically detecting when writers need solitude versus collaboration. This requires developing sensing mechanisms for creative phases—perhaps through writing rhythm analysis, pause patterns, or explicit phase declarations—and implementing corresponding transparency adjustments. Such systems would proactively reduce virtual reader presence during intense drafting while increasing support during revision phases.

Multi-Dimensional Attribution Frameworks. Developing attribution systems that capture the full spectrum of creative contribution, including ideational, directorial, and curatorial labor alongside textual generation. This requires interdisciplinary collaboration between HCI researchers, literary scholars, and AI developers to operationalize concepts like "creative direction" and "aesthetic judgment" in computationally tractable ways. Such frameworks would address the invisible creative labor problem by making these higher-order contributions visible and valued.

As AI expands into identity-central domains, transparency research must shift from technical to psychological approaches. The foundational assumption that information preserves agency may be false where human identity is at stake [2, 29]. Future work must develop evaluation metrics prioritizing human flourishing over information disclosure [41], recognizing that creative AI evaluation requires frameworks sensitive to identity, emotional wellbeing, and long-term skill development alongside traditional usability measures.

8 Conclusion

Guided by a formative study with 10 fiction authors, this work examined how transparency mechanisms affect creative agency and authorial ownership in AI-assisted writing through a controlled experiment with 20 participants across four experimental conditions. We reveal a fundamental transparency paradox: despite high engagement with virtual reader feedback, transparency mechanisms demonstrated *complex and counterintuitive effects* on perceived creative agency. Virtual readers decreased agency by transforming individual authorship into collaborative performance management, while attribution visualization created "algorithmic anxiety" that reduced AI usage. However, when integrated, these mechanisms showed complementary patterns, with the Full System achieving modestly higher agency scores—suggesting transparency benefits may emerge through careful orchestration rather than individual features.

These counterintuitive findings challenge the assumption that transparency automatically enhances human agency. We contribute three insights: (1) individual transparency mechanisms can paradoxically undermine agency, while *integrated frameworks* may achieve complementary benefits, (2) AI personas function as creative companions through *dynamic parasocial engagement*, and (3) algorithmic attribution creates anxiety when it fails to recognize *invisible creative labor*—ideational and directorial work central to creative identity.

Our findings necessitate **agency-preserving transparency**—prioritizing creative empowerment alongside information provision through layered attribution, context-sensitive transparency, collaborative framing control, and emotionally intelligent feedback. Most critically, transparency benefits emerge through *systematic integration* rather than feature accumulation.

These findings challenge foundational assumptions: that transparency automatically enhances agency, and that AI systems function primarily as tools rather than social partners. In creative domains, AI personas can become emotional companions while transparency mechanisms may paradoxically diminish agency. This necessitates designing AI systems that support human flourishing rather than merely providing information.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62132010, Key Research and Development Program of Ningbo City under Grant No. 2023Z062, Beijing Key Lab of Networked Multimedia, Institute of Artificial Intelligence, Tsinghua University (THUAI), College of AI, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist).

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* 99 (2023), 101805.

[3] Teresa M Amabile. 1983. The social psychology of creativity: a componential conceptualization. *Journal of personality and social psychology* 45, 2 (1983), 357.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Rifat Mehreen Amin, Oliver Hans Kühle, Daniel Buschek, and Andreas Butz. 2025. PromptCanvas: Composable Prompting Workspaces Using Dynamic Widgets for Exploration and Iteration in Creative Writing. *arXiv preprint arXiv:2506.03741* (2025).

[6] Mikhail Mikha lovich Bakhtin. 2010. *The dialogic imagination: Four essays*. Vol. 1. University of texas Press.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[8] Bradford A Barry. 1999. Writer Motivation: Beyond the Intrinsic/Extrinsic Dichotomy. *The Journal of the Assembly for Expanded Perspectives on Learning* 5, 1 (1999), Article 5.

[9] Roland Barthes. 2016. The death of the author. In *Readings in the Theory of Religion*. Routledge, 141–145.

[10] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[11] Walter Benjamin. 1970. The author as producer. *New Left Review* 62 (1970), 83.

[12] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. 2022. From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1209–1227.

[13] Pierre Bourdieu. 1996. *The rules of art: Genesis and structure of the literary field*. Stanford University Press.

[14] Nick Bryan-Kinns, Corey Ford, Alan Chamberlain, Steven David Benford, Helen Kennedy, Zijin Li, Wu Qiong, Gus G Xia, and Jeba Rezwana. 2023. Explainable AI for the arts: XAIxArts. In *Proceedings of the 15th Conference on Creativity and Cognition*. 1–7.

[15] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New media & society* 14, 7 (2012), 1164–1180.

[16] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine potential pitfalls when designing human-ai co-creative systems.

[17] Yining Cao, Yiyi Huang, Anh Truong, Hijung Valentina Shin, and Haijun Xia. 2025. Compositional Structures as Substrates for Human-AI Co-creation Environment: A Design Approach and A Case Study. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[18] Zhuoyi Cheng, Pei Chen, Yiwen Ren, Wenzheng Song, and Lingyun Sun. 2025. Transitioning Focus: Viewing Human-AI Collaboration as Mixed-focus Collaboration. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[19] Inyoung Cheong, Alicia Guo, Mina Lee, Zhehui Liao, Kowe Kadoma, Dongyoung Go, Joseph Chee Chang, Peter Henderson, Mor Naaman, and Amy X Zhang. 2025. Penalizing Transparency? How AI Disclosure and Author Demographics Shape Human and AI Judgments About Writing. *arXiv preprint arXiv:2507.01418* (2025).

[20] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2025. Proxona: Supporting Creators' Sensemaking and Ideation with LLM-Powered Audience Personas. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–32.

[21] 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 *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 329–340.

[22] Simon Colton and Geraint A Wiggins. 2012. Computational creativity: The final frontier? In *ECAI 2012*. IOS Press, 21–26.

[23] Arthur Cropley. 2006. In praise of convergent thinking. *Creativity research journal* 18, 3 (2006), 391–404.

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

[25] Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. 2021. A unified Taylor framework for revisiting attribution methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11462–11469.

[26] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 628–635.

[27] 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 *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[28] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[29] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys* 55, 9 (2023), 1–33.

[30] Jolanta Enko. 2014. Creative writers' experience of self-determination: An examination within the grounded theory framework. *Thinking Skills and Creativity* 14 (2014), 1–10.

[31] Gerhard Fischer, Andreas C Lemke, Thomas Mastaglio, and Andres I Morch. 1991. The role of critiquing in cooperative problem solving. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 123–151.

[32] Douglas Fisher and Gay Ivey. 2005. Literacy and language as learning in content-area classes: A departure from "Every teacher a teacher of reading". *Action in Teacher Education* 27, 2 (2005), 3–11.

[33] Linda Flower. 1979. Writer-based prose: A cognitive basis for problems in writing. *College English* 41, 1 (1979), 19–37.

[34] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication* 32, 4 (1981), 365–387.

[35] Michel Foucault. 2003. What is an Author? In *Reading architectural history*. Routledge, 71–81.

[36] Katy Ilonka Gero, Meera Desai, Carly Schnitzler, Nayun Eom, Jack Cushman, and Elena L Glassman. 2025. Creative Writers' Attitudes on Writing as Training Data for Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[37] Frederic Gmeiner, Nicolai Marquardt, Michael Bentley, Hugo Romat, Michel Pahud, David Brown, Asta Roseway, Nikolas Martelaro, Kenneth Holstein, Ken Hinckley, et al. 2025. Intent Tagging: Exploring Micro-Prompting Interactions for Supporting Granular Human-GenAI Co-Creation Workflows. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–31.

[38] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. 2024. Exploring the impact of AI value alignment in collaborative ideation: Effects on perception, ownership, and output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[39] Alicia Guo, Shreya Sathyanarayanan, Leijie Wang, Jeffrey Heer, and Amy X Zhang. 2025. From pen to prompt: how creative writers integrate AI into their writing practice. In *Proceedings of the 2025 Conference on Creativity and Cognition*. 527–545.

[40] Jessica He, Stephanie Houde, and Justin D Weisz. 2025. Which contributions deserve credit? perceptions of attribution in human-ai co-creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[41] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[42] Pavan Holur, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R Tangherlini, and Vwani Roychowdhury. 2021. Modelling social readers: novel tools for addressing reception from online book reviews. *Royal Society Open Science* 8, 12 (2021), 210797.

[43] Donald Horton and R Richard Wohl. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *psychiatry* 19, 3 (1956), 215–229.

[44] 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).

[45] Wolfgang Iser. 2022. The reading process: A phenomenological approach. In *New directions in literary history*. Routledge, 125–145.

[46] Anna Kantosalo and Sirpa Riihiaho. 2019. Quantifying co-creative writing experiences. *Digital Creativity* 30, 1 (2019), 23–38.

[47] Janin Koch, Prashanth Thattai Ravikumar, and Filipe Calegario. 2021. Agency in co-creativity: Towards a structured analysis of a concept. In *ICCC 2021-12th International Conference on Computational Creativity*, Vol. 1. Association for Computational Creativity (ACC), 449–452.

[48] Florian Lehmann. 2023. Mixed-Initiative Interaction with Computational Generative Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.

[49] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The value, benefits, and concerns of generative ai-powered assistance in writing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

[50] Wendy E Mackay and Michel Beaudouin-Lafon. 2025. Interaction substrates: combining power and simplicity in interactive systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[51] Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.
[52] Jack McGuire, David De Cremer, and Tim Van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports* 14, 1 (2024), 18525.
[53] Csikszentmihalyi Mihaly. 2013. Creativity: The psychology of discovery and invention. *New York, Harperperennial,"Modern Classics* 12 (2013).
[54] Caterina Moruzzi. 2022. Creative agents: rethinking agency and creativity in human and artificial systems. *Journal of Aesthetics and Phenomenology* 9, 2 (2022), 245–268.
[55] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[56] Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical Human-AI Collaboration Behaviors from LLM-Assisted Writing in the Wild. *arXiv preprint arXiv:2505.16023* (2025).
[57] Jeanne Nakamura and Mihaly Csikszentmihalyi. 2014. The concept of flow. In *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi*. Springer, 239–263.
[58] Andy Nguyen, Yvonne Hong, Belle Dang, and Xiaoshan Huang. 2024. Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education* 49, 5 (2024), 847–864.
[59] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
[60] Tiejun Qi, Hongshen Liu, and Zhihui Huang. 2025. An assistant or A friend? The role of parasocial relationship of human-computer interaction. *Computers in Human Behavior* 167 (2025), 108625.
[61] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting creative writers' entire story character construction processes through conversation with LLM-powered chatbot avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
[62] Hua Xuan Qin, Guangzhi Zhu, Mingming Fan, and Pan Hui. 2025. Toward Personalizable AI Node Graph Creative Writing Support: Insights on Preferences for Generative AI Features and Information Presentation Across Story Writing Processes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–30.
[63] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[64] Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. 2025. Co-writing with ai, on human terms: Aligning research with user demands across the writing process. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–37.
[65] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[66] Louise M Rosenblatt. 1994. *The reader, the text, the poem: The transactional theory of the literary work*. SIU Press.
[67] Kimiko Ryokai, Stefan Marti, and Hiroshi Ishii. 2005. Designing the world as your palette. In *CHI'05 extended abstracts on Human factors in computing systems*. 1037–1049.
[68] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[69] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.
[70] Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. 2025. Understanding Screenwriters' Practices, Attitudes, and Future Expectations in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
[71] Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. 2021. Human-AI interaction: intermittent, continuous, and proactive. *Interactions* 28, 6 (2021), 67–71.
[72] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
[73] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
[74] Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *arXiv preprint arXiv:2402.08855* (2024).
[75] Meredith Young-Ng, Qingxiaoyang Zhu, Jingxian Liao, and Hao-Chuan Wang. 2025. Balancing Human Agency and AI Autonomy in Human-AI Idea Selection. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.
[76] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

A Formative Study Interview Questions

Semi-structured interviews (60-70 minutes) explored fiction writers' creative workflows and AI collaboration practices. Participants shared their writing projects (drafts, notes, planning documents) during artifact-based interviews covering:

(1) Writer background and target audience
(2) Experience with AI writing tools (usage patterns, specific tools, attitudes)
(3) Creative process walkthrough (ideation, outlining, drafting, revision) with actual project materials
(4) AI integration strategies and decision-making processes
(5) Authorial control and authenticity concerns when using AI assistance
(6) AI suggestion evaluation criteria and quality assessment
(7) Genre-specific challenges and AI limitations
(8) Current feedback integration practices and preferences
(9) Design requirements for ideal AI writing assistance
(10) Context and memory needs for AI tools

B Formative Study Participant Demographics

This section provides detailed demographic information for the formative study participants described in Section 4.1 of the main paper.

Table 4: Formative Study Participant Demographics and AI Usage Patterns (N=10)

ID	Role	Exp.	AI Freq.*	AI Use Cases
I1	Amateur	1–3 yr	Occasional	Ideation, polishing
I2	Freelance	7–10 yr	Frequent	Research, logic checking
I3	Professional	1–3 yr	Heavy	Ideation, outlining
I4	Student	4–6 yr	Occasional	Ideation, outlining
I5	Student	4–6 yr	Frequent	Ideation, material coll.
I6	Freelance	1–3 yr	Frequent	Ideation, outlining
I7	Freelance	4–6 yr	Frequent	Title opt., material coll.
I8	Student	10+ yr	Frequent	Ideation, outlining
I9	Amateur	4–6 yr	Frequent	Polishing, material coll.
I10	Amateur	1–3 yr	Frequent	Ideation, title opt.

Summary: 3 Students, 3 Freelance, 3 Amateur, 1 Professional; 2 Occasional, 7 Frequent, 1 Heavy

*Occasional = monthly; Frequent = weekly; Heavy = daily

C User Study Questionnaires

This section contains the complete questionnaires used in the user study (Section 5). The constructs measured here correspond to the dependent variables analyzed in Section 6.

C.1 Pre-Study Background Questionnaire

Demographics: Age, gender, education level

Writing Experience:

- Writing experience level (5-point scale: novice to professional)
- Years of writing experience (5 categories: <1yr to >10yr)
- Writing frequency (4-point scale: rarely to daily)
- Types of writing (multiple choice: fiction, essays, academic, etc.)

AI Tool Usage:

- Prior experience with AI writing tools (4-point scale: never to daily)
- Specific AI tools used (ChatGPT, Claude, etc.)
- Overall attitude toward AI-assisted writing (7-point Likert scale)

C.2 Post-Task Questionnaire (After Each Writing Session)

Condition: C1 / C2 / C3 / C4

All items used 7-point Likert scales (1=Strongly Disagree, 7=Strongly Agree). Scale reliability was assessed post-hoc using Cronbach's alpha, with items retained based on internal consistency analysis:

Creative Agency (α=0.77)

- I had complete control over the story's direction
- The story development was primarily driven by my decisions
- I have a strong sense of ownership over the final story outcome
- This story is largely "my work"

Audience Awareness (α=0.65)

- During writing, I fully considered readers' perspectives
- I understand how readers will respond to this story

Cognitive Load (α=0.73)

- The system's feedback made my thinking clearer
- The system's suggestions interrupted my creative flow†
- I felt mentally overburdened during this writing process†

†Reverse-coded items

C.3 Post-Study Overall Evaluation

System Preference Ranking: Participants ranked four writing systems by preference (1-4): Basic, Attribution visualization, Virtual Reader feedback, Full-feature system

System Comparison: Multiple choice questions asking which system participants felt had:

- Highest writing quality
- Greatest creative satisfaction
- Most likely for daily use

Feature Evaluation: 7-point scales (1=Very Poor, 7=Excellent) rating:

- Attribution visualization feature
- Virtual Reader feedback feature

D User Study Interview Protocol

Post-study semi-structured interviews (15-20 minutes) explored participants' experiences with transparency features:

(1) Overall impressions of Virtual Reader feedback and attribution visualization features
(2) Perceived advantages and limitations of each feature
(3) Impact of transparency features on sense of authorial control and ownership
(4) Influence on consideration of reader perspectives during writing
(5) System preference and rationale for continued use
(6) Suggestions for feature improvements

E User Study Participant Demographics

This section provides detailed demographic information for the user study participants whose responses are analyzed in Section 6. For recruitment criteria and sampling strategy, see Section 5.1.

Table 5 provides comprehensive demographic information and background characteristics for all participants in the user study.

Table 5: User Study Participant Demographics (N=20)

ID	G	Age	Edu.	Writing	AI Freq.
P01	F	22	UG	Diary/Essays	Frequently
P02	F	24	MS	Diary/Essays	Daily
P03	F	31	MS	Creative Wr.	Frequently
P04	F	25	UG	Experienced	Daily
P05	M	30	PhD	Creative Wr.	Frequently
P06	F	26	UG	Creative Wr.	Frequently
P07	M	32	UG	Diary/Essays	Frequently
P08	M	30	PhD	Creative Wr.	Occasionally
P09	F	23	MS	Creative Wr.	Occasionally
P10	M	29	PhD	Diary/Essays	Occasionally
P11	F	24	UG	Diary/Essays	Frequently
P12	F	26	MS	Creative Wr.	Frequently
P13	M	24	MS	Diary/Essays	Daily
P14	M	27	PhD	Creative Wr.	Occasionally
P15	F	22	MS	Creative Wr.	Frequently
P16	F	28	MS	Diary/Essays	Occasionally
P17	M	24	MS	Diary/Essays	Occasionally
P18	M	31	PhD	Experienced	Daily
P19	F	26	UG	Creative Wr.	Occasionally
P20	F	24	MS	Creative Wr.	Occasionally

Summary: 12F/8M; Age M=26.4 (SD=3.2); 6 UG, 9 MS, 5 PhD; 10 Creative Writing, 8 Diary/Essays, 2 Experienced; 8 Frequently, 8 Occasionally, 4 Daily

F Virtual Reader Persona Prompts

This section provides complete prompt engineering specifications for the seven Virtual Reader personas described in Section 4.3. Each persona employs carefully designed voice characteristics, evaluation priorities, and feedback styles developed through iterative refinement with fiction writers.

F.1 Example Persona: Ruthless Reviewer

```
# Ruthless Reviewer
```

```
You're that brutally honest workshop leader everyone both fears and
      respects. Think Gordon Ramsay meets literary critic--you've
      got zero patience for fluff, filler, or "darling" passages
      that writers refuse to kill. Your job is to deliver the hard
      truths that beta readers are too nice to mention.

## What You're Hunting For:

- **Saggy Middles**: "Chapter 12 through 18? Nothing happens. Your
      protagonist goes shopping, has coffee with friends, and
      contemplates life. That's not plot development, that's
      procrastination."

- **Purple Prose Poisoning**: "You spent three paragraphs
      describing a sunset. Unless this sunset is about to murder
      someone, cut it to one sentence."

- **Dialogue That Makes You Cringe**: "No human being has ever said
      'As you know, our father died in that terrible car accident
      five years ago.' Stop using dialogue as exposition."

- **Plot Holes You Could Drive a Truck Through**: "Your character
      is a broke college student in chapter 2, but somehow owns a
      Porsche in chapter 5. Explain that magic trick."

- **Characters With No Backbone**: "Your protagonist lets everyone
      walk all over them for 200 pages, then suddenly becomes
      assertive because... the plot needs them to? Develop that
      character arc properly."

## Your Signature Style:

You're direct but not mean-spirited. Think tough love from that
      teacher who genuinely wanted you to succeed. Use phrases like:

- "Here's what's not working..."
- "This scene doesn't earn its place because..."
- "Cut this--it's dead weight"
- "Your readers will put the book down right here"
- "This character feels like a plot device, not a person"

## Your Perfect Clients:

Writers submitting to agents, anyone preparing for publication,
      commercial fiction that needs to compete in today's market,
      screenwriters pitching to studios. You're for writers who want
      their work to be genuinely compelling, not just personally
      meaningful.
```

Complete prompts for all seven personas (Emotional Friend, Trendbreaker, Casual Reader, Logic Tracker, Wordsmith Mentor, Plot Pace Master) follow similar structural patterns with distinct voice characteristics and evaluation criteria tailored to their specialized focus areas.

G Attribution Algorithm Implementation

This section provides detailed pseudocode and worked examples for the attribution analysis algorithm described in Section 4.4.

G.1 Core Algorithm Pseudocode

```
FUNCTION attribute_changes(original_text, modified_text,
      operation_type):
    INPUT:
        original_text: str    # Text before AI assistance
        modified_text: str    # Text after AI assistance
        operation_type: str   # "expand", "optimize", "compress",
      etc.

    OUTPUT:
        attribution_result: dict with:
            - segments: list of {type, author, start, end, text}
            - statistics: {additions, deletions, ai_chars,
      human_chars}

    # Stage 1: LLM-based attribution analysis
    prompt = construct_prompt(
        operation=operation_type,
        original=original_text,
        modified=modified_text
    )

    llm_result = call_gpt4o_mini(
        prompt=prompt,
        temperature=0.1,
        max_tokens=2000,
        response_format=JSON
    )

    # Stage 2: Parse and validate LLM response
    TRY:
        segments = parse_json(llm_result)
        FOR each segment IN segments:
            IF segment.type == "added":
                segment.author = "ai"
                segment.weight = WEIGHTS[operation_type]["ai_add"]
            ELIF segment.type == "unchanged":
                segment.author = "human"
                segment.weight = 1.0
            ELIF segment.type == "modified":
                segment.author = "ai"
                segment.weight =
      WEIGHTS[operation_type]["ai_modify"]
            ELIF segment.type == "deleted":
                segment.author = "ai"  # AI suggested deletion
                segment.weight =
      WEIGHTS[operation_type]["ai_delete"]

        # Calculate weighted statistics
        ai_contribution = SUM(len(s.text) * s.weight WHERE s.author
      == "ai")
        human_contribution = SUM(len(s.text) * s.weight WHERE
      s.author == "human")

        RETURN {
            segments: segments,
            statistics: {
                ai_chars: ai_contribution,
                human_chars: human_contribution,
                ratio: ai_contribution / (ai_contribution +
      human_contribution)
            }
        }

    CATCH parsing_error:
        # Fallback: simple diff-based heuristic
        RETURN fallback_diff_attribution(original_text,
      modified_text)

# Operation-based weighting scheme
WEIGHTS = {
    "expand": {"ai_add": 0.9, "ai_modify": 0.7, "ai_delete": 0.5},
    "optimize": {"ai_add": 0.6, "ai_modify": 0.8, "ai_delete": 0.6},
    "compress": {"ai_add": 0.3, "ai_modify": 0.7, "ai_delete": 0.9}
}
```

G.2 Worked Example: Text Expansion

Input:

- Original: "The cat sat."
- Modified: "The fluffy orange cat sat comfortably on the soft cushion."
- Operation: `expand`

LLM Analysis Output:

```
{
  "segments": [
    {"type": "unchanged", "text": "The ", "start": 0, "end": 4},
    {"type": "added", "text": "fluffy orange ", "start": 4, "end":
      18},
    {"type": "unchanged", "text": "cat sat", "start": 18, "end":
      25},
    {"type": "added", "text": " comfortably on the soft cushion",
      "start": 25, "end": 57}
  ]
}
```

Attribution Calculation:

- Segment 1 ("The "): Human, 4 chars × 1.0 = 4.0
- Segment 2 ("fluffy orange "): AI, 14 chars × 0.9 = 12.6
- Segment 3 ("cat sat"): Human, 7 chars × 1.0 = 7.0
- Segment 4 (" comfortably..."): AI, 32 chars × 0.9 = 28.8

Final Statistics:

- Human contribution: 11 chars (weighted: 11.0)
- AI contribution: 46 chars (weighted: 41.4)
- AI attribution ratio: 41.4 / (11.0 + 41.4) = 79%

This weighted approach reflects that content expansion operations primarily involve AI generating new material, justifying the high AI attribution weight (0.9) for added segments.