

DraftMarks: Enhancing Transparency in Human-AI Co-Writing Through Interactive Skeuomorphic Process Traces

Momin N. Siddiqui

Georgia Institute of Technology
USA
msiddiqui66@gatech.edu

Nikki Nasser

University of California, Berkeley
USA
nassern1@berkeley.edu

Adam Coscia

Georgia Institute of Technology
Atlanta, Georgia, USA
acoscia6@gatech.edu

Roy Pea

Stanford University
USA
roypea@stanford.edu

Hari Subramonyam

Stanford University
USA
harihars@stanford.edu

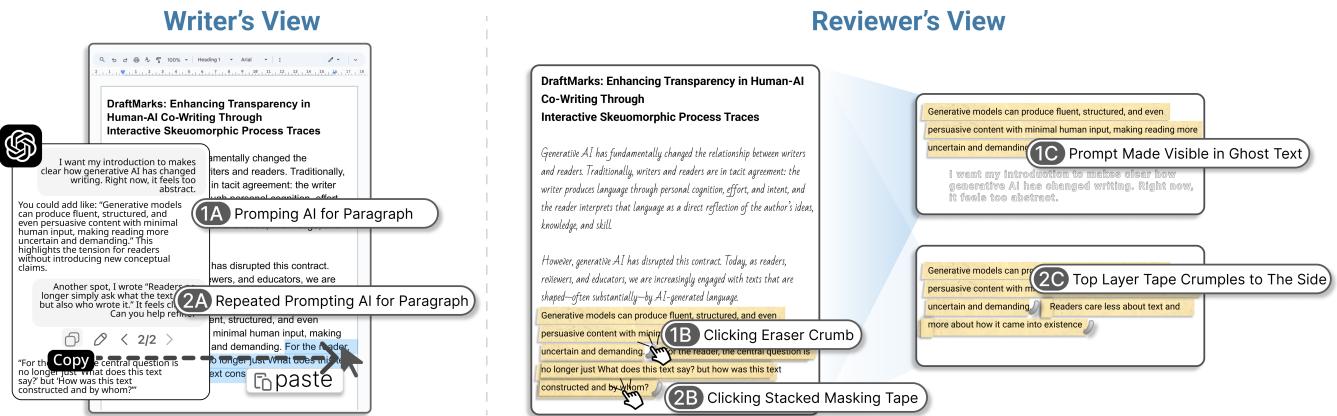


Figure 1: DraftMarks reveals the hidden dynamics of human-AI collaboration to readers. The Writer's View shows AI generation calls: single prompts (1A) and repeated prompting (2A). The Reviewer's View surfaces these interactions within the final text through skeuomorphic design: eraser crumb and ghost text (1B-1C), and stacked masking tape (2B-2C).

Abstract

As generative AI becomes part of everyday writing, questions of transparency and productive human effort are increasingly important. Educators, reviewers, and readers want to understand how AI shaped the process. Where was human effort focused? What role did AI play in the creation of the work? How did the interaction unfold? Existing approaches often reduce these dynamics to summary metrics or simplified provenance. We introduce DraftMarks, an augmented reading tool that surfaces the human-AI writing process through familiar physical metaphors. DraftMarks employs skeuomorphic encodings such as eraser crumbs to convey

the intensity of revision, and masking tape or smudges to mark AI-generated content, simulating the process within the final written artifact. By using data from writer-AI interactions, DraftMarks' algorithm computes various collaboration metrics and writing traces. Through a formative study, we identified computational logic for different readership, and evaluated DraftMarks for its effectiveness in assessing AI co-authored writing.

Keywords

User interface, Human-AI collaboration, Writing analytics, Skeuomorphism, Large language models

ACM Reference Format:

Momin N. Siddiqui, Nikki Nasser, Adam Coscia, Roy Pea, and Hari Subramonyam. 2018. DraftMarks: Enhancing Transparency in Human-AI Co-Writing Through Interactive Skeuomorphic Process Traces. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 21 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 Introduction

Generative AI has fundamentally changed the relationship between writers and readers. Traditionally, writers and readers are in tacit

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY'

© 2018 ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXX.XXXXXXX>

agreement: the writer produces language through *personal cognition*, effort, and intent, and the reader interprets that language as a direct reflection of the author’s ideas, knowledge, and skill [33, 62]. Whether reading an essay, an email, or a scientific article such as this article, the reader makes inferences about the author’s expertise, creativity, and trustworthiness based on the text itself [49]. However, generative AI has disrupted this contract. Today, as readers, reviewers, and educators, we are increasingly engaged with texts that are shaped—*often substantially*—by AI-generated language. Generative models can produce fluent, structured, and even persuasive content with minimal human input, making reading more uncertain and demanding. For the readers, the central question is no longer just *What does this text say?* but also *How was this text constructed and by whom?*

Unfortunately, readers lack effective tools and transparent design affordances to make these distinctions and critically assess texts co-authored with AI as a writing partner. Current approaches are primarily limited to AI usage disclosures [18], sharing AI prompts and conversation logs [83], and token-level attribution maps or summary visualizations of AI contributions [34, 64]. However, these approaches are often external to the text rather than *within* it, requiring the reader to step out of the reading experience to consult metadata or interpret visualizations detached from the comprehension flow. This separation becomes problematic when readers are making local interpretive decisions: *Is this sentence the author’s reasoning or the model’s default phrasing? Was this claim carefully revised or copied verbatim?* For example, knowing that a paragraph was iteratively revised might lead a reader to trust its reasoning more; recognizing that a definition was injected via AI may prompt the reader to cross-check its accuracy. These decisions are part of how the readers build mental models about what matters in writing, what to question critically, and what reflects human expertise [63]. When process information is abstracted away, the reader’s comprehension remains incomplete, obscuring the *cognitive labor* — or lack thereof — behind the text.

In this work, we explore the use of skeuomorphic design [7, 19, 53], i.e., visual encodings grounded in familiar physical metaphors, to *surface* human-AI writing interactions in context. Rather than abstract icons or disconnected dashboards, skeuomorphic cues such as masking tape, eraser crumbs, smudges, and scribbled margins could provide legible embedded markers of writing activity. Akin to traditional settings, readers may draw meaning about authorial effort from crossed-out sentences, margin notes, use of white-out tape, or layered corrections of pen over pencil (i.e., physical traces into the writer’s thinking). Readers rely on physical signals to support *inline* interpretation of AI coauthored text. These signals should indicate where the writer invested effort, where they accepted or rejected AI suggestions, and where the text evolved through iteration. For example, a masking tape overlay might indicate a passage written primarily by an AI model, while eraser marks could show a region of intensive human-AI iteration. These metaphor-driven encodings draw on readers’ pre-existing literacies for interpreting physical documents and editing marks, repurposing them for AI-mediated writing.

In our technique, we formalize this approach by linking interaction logs of human-AI writing to skeuomorphic visualizations that

appear within the text document. DraftMarks follows a model-view-controller paradigm, where the model processes interaction data to compute metrics such as contribution ratios, edit frequency, and prompt iteration depth, and the view maps them to visual overlays that retain the spatial and rhetorical coherence of the original text. To devise the controller logic that determines how and when these visualizations appear, we conducted a formative study with 21 readers across three readership roles (teachers, academic reviewers, and general readers), gathering insights on their interpretation needs and preferences for process transparency. We found that while teachers prefer detailed traces to support formative assessment of student learning, academic reviewers prefer minimal process information that prioritizes intellectual contribution over process details. General readers seek authenticity markers to evaluate author effort and trustworthiness. These findings informed our controller design, which adapts the same underlying interaction data into stakeholder-specific visualizations that balance comprehension support with evaluative needs.

As shown in Figure 1, the human-AI collaboration trace (1A and 2A) is rendered using DraftMarks to surface co-writing insights directly into the text. In case of the reviewer view, we see single masking tape and its variant of stacked masking tape, encoding single prompt and iterative prompting respectively. Together, these features in DraftMarks enable *process-aware reading* by offering *transparency* into how the text was constructed with AI. A between-subject evaluation with 70 participants split between DraftMarks and a baseline comprising of collaboration chat history and final essay artifact in split view revealed that participants in DraftMarks condition had better comprehension of writing process and also reported higher self-reported transparency in the co-writing process.

In summary, our contributions are:

- DraftMarks — a visualization technique that uses skeuomorphic design to embed writing process signals directly within AI-coauthored text.
- A tool implementation following the model-view-controller paradigm that takes human-AI collaboration data from writing sessions and automatically generates DraftMarks visualizations, mapping interaction data to in-text visual cues.
- A design probe with 21 readers across 3 reader types (educators, reviewers, and general readers) that informed the design of our controller logic and revealed key requirements for process transparency in AI-coauthored text.
- A user evaluation showing that DraftMarks supports more accurate and transparent interpretations of the co-writing process than a baseline disclosure condition, while maintaining usability and low cognitive load.

2 Related Work

AI co-authorship, the practice of collaboratively producing text with large language models, has introduced a fundamental change in how we think about writing, authorship, and intellectual ownership [41, 42]. These “co-creative systems” operate along a continuum between user and system contributions, blurring clear lines of attribution and complicating questions of responsibility and credit [13]. Academic institutions and publishers have responded

with caution: top-tier journals, including *Nature* and *Science*, prohibit listing AI as an author and require clear disclosure of AI participation [8, 30, 66]. However, cases of AI systems such as ChatGPT [55] being listed as authors have already appeared in academic databases [3].

Central to these concerns is a question of accountability. Tools such as ChatGPT “cannot be held responsible” explaining why editors and publishers reject AI coauthorship outright [56]. Furthermore, AI tools introduce an “AI Ghostwriter Effect” where users do not perceive ownership of AI-generated text, but also refrain from publicly acknowledging AI’s role [21]. These tensions are particularly acute in educational settings, where learning to write is as important (if not more so) than the final product. Several studies note that reliance on AI-generated writing can erode students’ critical thinking, expressive capabilities, and revision skills [58]. Teachers often express concern that students will bypass the cognitive work of structuring ideas or developing arguments, a shortcut that undermines learning goals and the validity of writing as an assessment [45]. Moreover, the growing presence of fluent but untraceable AI-generated prose makes it difficult for instructors to distinguish genuine student effort from automated composition [38, 51]. Similar concerns arise in peer review journal article submissions: reviewers may unknowingly assess work with extensive AI involvement, yet be unable to evaluate the quality of argumentation, originality, or authorial voice [49].

To address this gap between private awareness and public disclosure, our work examines how visualizing the writing process can support more transparent, reflective, and interpretable human-AI collaboration. This section reviews previous work on LLM writing support tools, visualization systems to understand human-AI interaction, and text visualization techniques to analyze collaborative writing to motivate the design of *DraftMarks*.

2.1 LLM Writing Support Tools

The ubiquity of LLM conversational interfaces has spawned a new paradigm of AI-assisted writing. Recent LLM-driven writing tools are increasingly aiming for a more synergistic human-AI partnership. This development is consistent with Douglas Engelbart’s advocacy of the idea of augmenting human intellect through the coevolution of human and tool systems [22, 23]. He argued that humanity’s ability to solve problems improves as our “tool systems” (our technology) and “human systems” (our organization, language, and methods) evolve together in a positive feedback loop. Tools such as Script&Shift [65], VISAR [84], and CoAuthor [42] focus on enabling writers to meaningfully integrate AI contributions within their own workflow, fostering a more interactive and collaborative process rather than the simple acceptance of suggestions.

Critically, most previous tools are mixed-initiative [84], in which the LLM often takes a proactive role in influencing or even driving the writing process. However, recent studies have raised ethical concerns about AI-assisted writing. The use of AI assistance in writing raises significant concerns about decreased author agency and ownership [50], the generation of lackluster or stereotyped content [14, 27], and ethical issues, including plagiarism and the

possibility for the views of the user of a tool to be unduly influenced [37, 80]. Although maintaining some mixed-initiative capabilities, AI-assisted writing tools often lend themselves to AI heavily influencing or even driving the writing process, potentially reducing the human to a more passive role [50], thus simply a ‘human in the loop’ rather than a ‘human in the center of the loop’. In addition, specific concerns have been raised about the use of AI by students in academic writing [9, 20, 24, 49].

To address these concerns, a critical question persists: How might we enable transparency in this collaborative dynamic? There is a pressing need to provide writers, educators, reviewers, and other stakeholders with clear and informative data and associated tools to enable detailing how AI was used during the writing process. As highlighted in Hoque et al. [34], tools are needed to help writers track their agency and AI usage, particularly as institutions and publishers establish policies that require transparency in AI-assisted writing [8, 57]. This need for transparency in the process and insightful analytics motivates our development of *DraftMarks*.

2.2 Visual Analytics For Human-AI Collaboration

Collaborative writing has a long history of research into how humans coordinate with other humans to accomplish writing tasks [4]. We draw inspiration from prior works that visualize contributions from multiple authors towards analyzing writing patterns. Many prior tools focus on “distant reading,” revealing aggregate patterns, structures, authorship contributions, or timelines without showing the fine-grained evolution of the text itself [15, 40, 47, 75, 78]. For example, DocuViz [78] is a tool for analyzing human-human collaborative writing post hoc. These visualizations can be effective for understanding contribution ratios or identifying cooperation and conflict patterns, but are often difficult for laypeople to interpret, and can obscure the close reading of content changes [75]. Techniques like highlighting authorship [17, 34, 71] struggle to represent the complex history of revisions, especially when content ownership changes. In contrast, fewer tools support “close reading” of collaborative revisions. Some approaches use text animations or highlighting within a timeline view to show how content changes over time [17]. Other efforts have explored sentence-level visualizations [64], but these tend toward abstract encodings that hinder close reading and may not capture metrics such as iteration depth or the nature of writer engagement. Although useful, these tools often focus mainly on the *what* of the substance of the changes, not necessarily on the *how* of the efforts of the writing processes underlying text production.

Developing visual analytics systems for human-AI collaboration is a relatively nascent but growing area of research. Wang et al. describe the characteristics of systems that foster human-AI collaboration, which can be explored and analyzed using visual analytics systems [77]. Rogers and Crisan develop a methodology for visualizing collaborative human-AI processes in data work [61]. Li et al. characterize the collaborative roles of humans in a human-AI storytelling workflow, which can be visualized [44]. Several domains have been the focus of visual analytics systems that facilitate human-AI collaboration, mainly in the area of interactive data analysis [67]. For example, StuGPTViz is a visual analytics

system for post hoc analysis of student-LLM collaboration in education tasks [16]. Graphologue [39] and Sensecape [68] are systems that facilitate interactive human-AI collaboration on sensemaking tasks. There is a recognized need for visualizations that specifically illuminate critical engagement with AI suggestions across domains [63].

Yet while there is prior work on visualizing human-AI collaboration, visual analytics tools specifically for gaining insights into human-AI collaborative writing are almost nonexistent. However, HaLLMark [34] is a visual analytics system that represents a significant step, offering interactive visualizations to support provenance tracking and transparency regarding AI use, with the aim of helping writers maintain agency and comply with policies [8, 8, 54]. It captures key metrics such as prompt categories and AI response integration. However, the authors acknowledge its limitations for analyzing long-form writing and suggest future work towards platform-independent tracking. Crucially, HaLLMark focuses primarily on provenance (what text came from where) and summary statistics, offering less insight into the productive struggle or iterative refinement process undertaken by the writer, aspects often of interest to readers, educators, and reviewers. The productive struggle of a learner is of special interest to educators, as it signals the participation of cognitive processes associated with advances in learning [6]. Thus, we identify a critical gap: the need for visualizations that go beyond simple contribution ratios and provenance specifically in writing tasks. DraftMarks aims to fill this gap by visualizing more complex process metrics reflecting the writer's effort, iteration, and metacognitive engagement with the AI, directly within the context of the final document.

2.3 Text Visualization Techniques

Text visualization encompasses a wide range of techniques for representing textual data [10, 74, 76, 79, 81]. Common methods include *word clouds* to summarize the frequency of words [74] and *tree structures* to show concordance [79]. Many techniques focus on "distant reading," analyzing large corpora to reveal patterns, topics, and relationships through spatial layouts, network diagrams, or temporal flows [25, 73]. Other techniques support "close reading" by overlaying information on the text itself, such as highlighting phrases based on the underlying data or tags [17]. Recent work has also explored using AI and visualization to analyze specific textual features, such as character representation or sonic properties in poetry [52], or to visualize the structure of LLM responses themselves [39, 68]. Our work draws on this rich history, but focuses specifically on embedding visualizations of the writing process *within* the document to support interpretation of human-AI collaboration during close reading. This strategy necessitates moving beyond standard text visualization techniques towards metaphors that convey process history, leading us to explore skeuomorphism, as we describe in the next section.

3 User Scenario

At a high level, DraftMarks is a web-based *interactive text viewer* designed to make the invisible dynamics of human-AI writing legible to readers by embedding process cues directly into the text. As

shown in Figure 2, the interface overlays skeuomorphic encodings—visual marks inspired by everyday writing materials—to reveal how a piece of text came to be. The augmentations, including masking tape, eraser crumbs, smudge marks, ghost text, stencil marks, and residual glue, were selected for their familiarity and semantic fit with the underlying activity they represent, such as provenance, erasure, or absence (discussed in more detail in Section 5). This visual language transforms the final draft into a layered record of interaction, preserving readability while revealing the writing processes.

To understand the user experience of DraftMarks, let us consider an example scenario of Miss Jones, an eighth-grade English teacher. In her classroom, students are increasingly experimenting with generative AI to scaffold their narrative writing. Miss Jones has established a clear classroom policy: "*You are allowed to use AI for brainstorming and improving tone. However, you must generate your own original ideas and may not use AI to write complete paragraphs.*" Ordinarily, Miss Jones would ask students to share their essays alongside their ChatGPT transcripts to check compliance with her policy. Today, she decided to try DraftMarks. In the following, we walk through her analysis of three students' essays (Fig. 2), Matilda, Lavender, and Bruce.

3.1 Reviewing Matilda's Essay: Using AI for Transitions & Transformation

When Miss Jones opens Matilda's essay (Fig 2A), she immediately notices several visual encodings (Fig 3A), signaling AI involvement at different points. Miss Jones prefers to read in two passes: first, to comprehend Matilda's writing, then to assess it against her rubric and prepare feedback. On her first pass, she notes smudge marks in the first paragraph and masking tape at the beginning of paragraphs two and three. Both represent AI-generated text, but with different functions: the masking tape indicates inserted content, while smudge marks show AI tone transformations of existing text. The masking tape metaphor suggests something provisional or externally added, and the 'crumpled' state further conveys whether the suggestions were kept intact, modified, or fragmented. This supports Miss Jones's rapid inference about the degree of human intervention. In her second pass, Miss Jones examines the crumpled masking tape in the second paragraph (Fig. 3B), indicating that Matilda deleted words within an AI insertion: "*Moving from fear to understanding why I felt this way requires looking back to when it started.*" Clicking the tape reveals the original AI text for five seconds before it crumples again. Next, Miss Jones clicks on the eraser crumb beside this tape (Fig. 3C), revealing Matilda's ghost text prompt requesting help with a transition from the opening paragraph (Fig. 3D). She finds a similar transition request related to the masking tape in the third paragraph.

At the start of the essay, multiple residual glue marks indicate attempts to generate full drafts with AI that Matilda later discarded. Clicking on one (Fig. 3E) shows the rejected AI text (Fig. 3F). Miss Jones infers that Matilda rejected the AI draft in favor of her own writing. Finally, she examines the smudge marks in the first paragraph. Since these encode tone transformations of existing content, Miss Jones decides not to click into the underlying prompts. From these traces, she concludes that: *Matilda struggles with transitions,*

 " Write a story about a character who has to overcome a fear or a challenge that seems impossible at first.
Your story should show how the character changes from the beginning to the end."

A Matilda 

The community pool's diving board stretched out like a plank over churning blue waters, and I stood frozen at its edge. My harsh breathing echoed in my ears as other kids splashed and laughed below. I had been coming to this pool every summer since I was six, but that diving board might as well have been Mount Everest.

Moving from fear to understanding why I felt this way requires looking back to when it started. When I was seven, my cousin Jake had pushed me off the diving board as a joke. I remember the shock of hitting the water wrong, the burning chlorine rushing up my nose, and the terrifying moment when I couldn't tell which way was up. Ever since then, heights over water made my stomach flip.

The next few seconds would determine whether I remained trapped by my fear or finally broke free from it. I closed my eyes, took three deep breaths, and before I could talk myself out of it, I jumped.

The fall felt endless. Water rushed toward me, and for a split second, that old panic clawed at my chest. But this time was different. This time, I was ready. I hit the water cleanly, sank down into the cool depths, and pushed off the bottom with my legs. When I surfaced, gasping and grinning, Maya was cheering so loud that other swimmers turned to look.

"I did it!" I shouted, treading water and feeling lighter than I had in years. That day changed something fundamental in me. I realized that courage isn't about not being scared—it's about being scared and doing it anyway. The diving board taught me that some fears are just old stories we keep telling ourselves, and sometimes we have to write a new ending.

B Lavendar 

Public speaking had always been my worst nightmare, so when Mrs. Peterson announced our final presentation would be worth thirty percent of our grade, I felt my world crashing down around me.

I sat in my bedroom that night, staring at my notes and feeling completely overwhelmed by the task ahead of me.

Explain choice The presentation was supposed to be about a historical figure we admired, and I had chosen Marie Curie because her story of perseverance inspired me. But how could I talk about someone so brave when I couldn't even raise my hand in class?

Add detail For weeks, I practiced in front of my mirror, stumbling over words and watching my face turn red. My mom suggested I practice with our dog, Buster, but even he seemed to judge my shaky voice. I tried recording myself, but hearing my own voice made everything worse.

The night before the presentation, I called my older sister Emma for advice. "What if I mess up?" I asked her, my voice barely above a whisper.

"Then you mess up," she said simply. "But what if you don't?" Standing in front of the classroom the next morning, I felt like I might faint. Twenty-five pairs of eyes stared at me expectantly. My hands shook as I held my notecards, and my prepared opening line disappeared from my memory completely.

Modify tone But then I looked at my poster of Marie Curie and remembered why I chose her. Here was a woman who faced rejection, discrimination,

C Bruce 

Last summer, I faced the most daunting challenge of my life when my family decided to embark on a hiking expedition to summit Mount Washington, the highest peak in the northeastern United States. As someone who had never engaged in serious outdoor activities beyond casual neighborhood walks, the prospect of conquering this formidable mountain seemed utterly impossible, filling me with a mixture of anxiety and determination that would ultimately transform my understanding of personal resilience and capability.

The preparation phase proved to be equally challenging as the actual climb itself, requiring weeks of intensive physical conditioning, equipment acquisition, and mental preparation that pushed me far beyond my comfort zone. My parents, both experienced hikers, insisted on a rigorous training regimen that included daily runs, weekend practice hikes on local trails, and strength-building exercises that left me exhausted but gradually more confident in my physical abilities.

The morning of our departure arrived with crisp mountain air and clear skies that seemed to promise a perfect day for our adventure. My backpack felt heavy with supplies as we began our ascent on the Tuckerman Ravine Trail. The first mile was manageable, winding through dense forest with dappled sunlight filtering through the canopy above us.

But as we climbed higher, the trail became increasingly steep and rocky, testing every muscle in my legs and challenging my cardiovascular endurance in ways I had never experienced before.

Figure 2: Human-AI Co-Writing Essay Samples.

but is actively working to improve them, using AI as a scaffold rather than a crutch. Her feedback encourages Matilda to experiment with drafting her own transitions before refining them with AI.

3.2 Reviewing Lavender's Essay: Using AI for Iterative Generation & Feedback

Lavender's essay illustrates a different pattern (Fig 2B). During her first pass, Miss Jones notices residual glue, masking tape, eraser crumbs, and stencil marks. Although few AI sentences remain in the final draft, Lavender appears to have leaned heavily on AI throughout the process. At the end of the first paragraph, a residual glue mark (Fig 4D) reveals an AI-generated outline (Fig 4E), ultimately deleted but clearly echoed in the final structure of Lavender's essay. The second paragraph (Fig 4A) contains a masking tape with two eraser crumbs, one darker than the other. The darker crumb signals a more complex prompt. Clicking on it reveals one sentence in ghost text and another in masking tape (Fig 4B). Because Lavender included a previously generated AI sentence in her prompt, this nesting creates a chain of iterations (Fig 4A–C). Over the taped sentence "But as I walked home, I remembered what my grandmother had whispered to me before she died—and suddenly, I knew exactly what to do.", Miss Jones also sees segmented gray smudge marks, indicating recursive tone adjustments (Fig 4B). Clicking on the associated crumb confirms that the AI was repeatedly asked to refine its own text (Fig 4C).

In addition, the third, fourth, and sixth paragraphs show stencil marks: solid in the third and fourth, hollow in the sixth (Fig 2). These encode AI feedback requests, with solid marks showing integrated changes and the hollow mark showing rejected feedback. For example, in paragraph three, clicking the stencil mark (Fig 4F)

reveals the AI feedback that helped Lavender expand her reasoning for choosing Marie Curie (Fig 4G). From this move, Miss Jones concludes: *Lavender is doubtful about her writing*, seeking reassurance through repeated generation and critique. This is less about laziness than about uncertainty: she is looking for confidence in her authorial voice. In her feedback, Miss Jones highlights moments where Lavender's own drafts were strong without AI assistance, encouraging her to trust her instincts more than her writing process revealed that she did.

3.3 Bruce: Using AI for Passive Writing

Bruce's essay shows yet another pattern (Fig 2C). When Miss Jones opens it, she is struck by the overwhelming presence of masking tape. Two full paragraphs are entirely taped, which means that they were fully AI-written, while the third paragraph is only partially taped, with gaps showing Bruce's own contributions. Throughout the essay, there is only a single eraser crumb at the end of the fourth paragraph. Its faint color suggests a simple, low-effort prompt. Clicking on it (Fig 4I) reveals ghost text showing that Bruce had pasted the entire class assignment into ChatGPT (Fig 4J), generating an essay with minimal input beyond one short addition in paragraph three.

Unlike Matilda's selective scaffolding or Lavender's iterative exploration, Bruce's work appears largely AI-written, with little evidence of his own thinking or revision. Although the essay looks polished, the skeuomorphic cues expose its shallow authorship. In her feedback, *Miss Jones notes the lack of original writing and stresses the importance of Bruce developing his independent writing skills. She warns that overreliance on AI not only hinders his growth but also violates her classroom policy.*

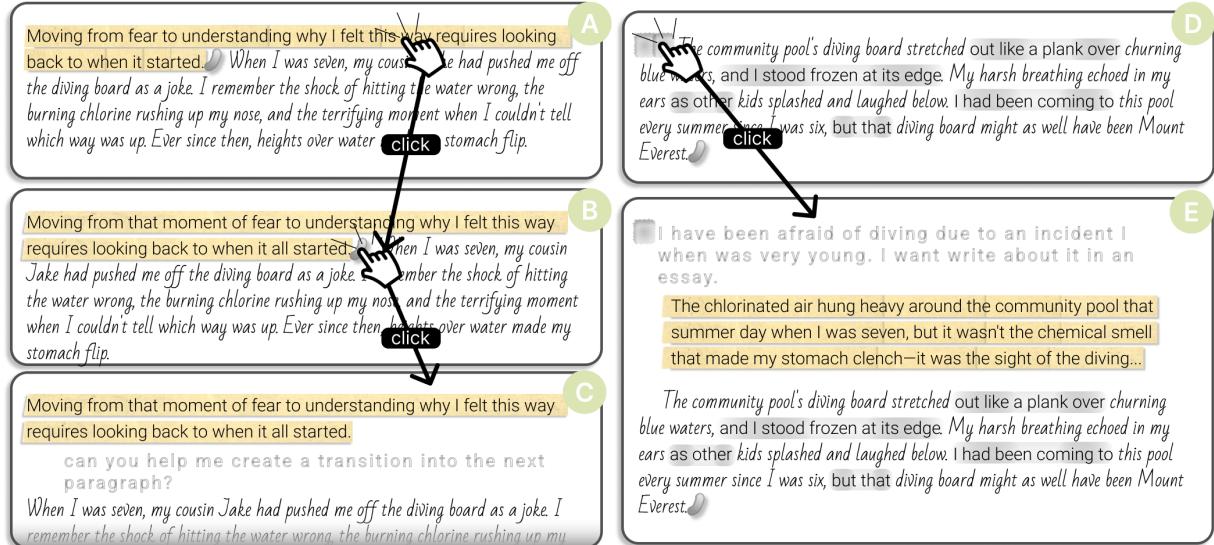


Figure 3: Usage scenario exploration of 2A Matilda' essay.

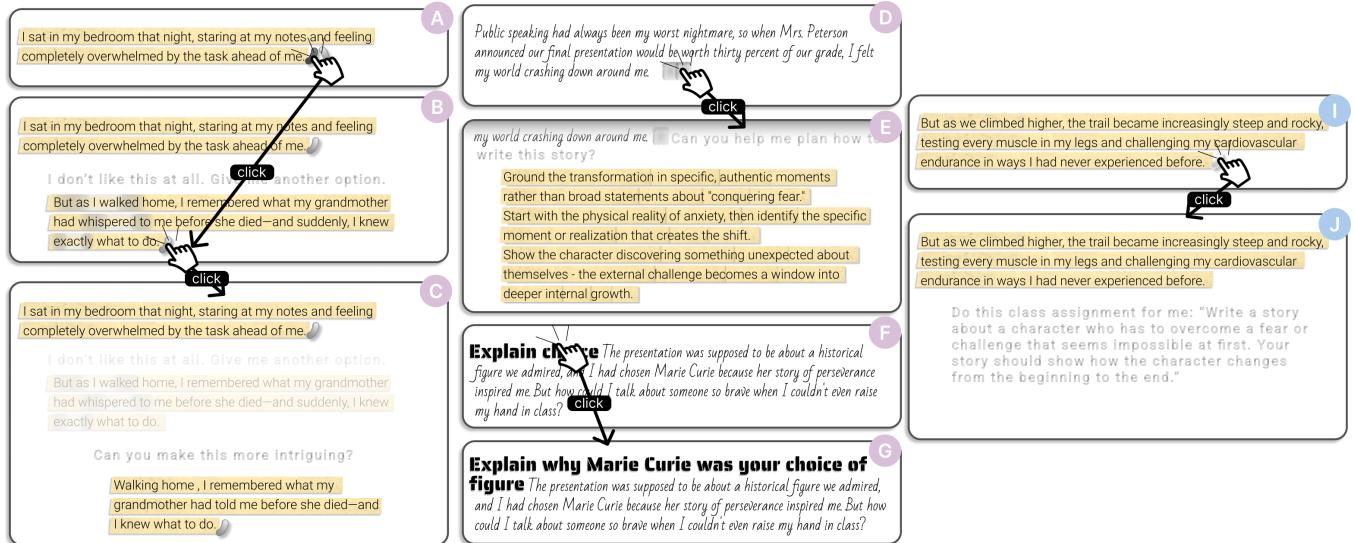


Figure 4: Usage scenario exploration of 2B Lavendar and 2C Bruce's essay.

Taken together, these three cases illustrate how DraftMarks makes different modes of AI-assisted writing visible. Matilda uses AI selectively as a scaffold, Lavender iterates recursively to build confidence, and Bruce relies passively on AI for entire sections. By revealing these distinctions, DraftMarks allows teachers like Miss Jones to tailor the feedback to the needs of each student, something that would be impossible with plain text alone.

4 DraftMarks' Data Model

As shown in Figure 5, DraftMarks is implemented as a model-view-controller (MVC) architecture. Here we elaborate on the data model

for DraftMarks (Figure 5a), and in (Sections 5 and 6) we discuss the other components.

Our system captures human-AI collaborative writing through version-controlled editor states that preserve the complete interaction history. In most modern rich text editors [29, 32, 48, 70], each document is stored as a hierarchical node-based structure, often called an editor state, as shown in figure 6 A. Nodes can be structural elements such as paragraphs, headings, lists, or text nodes that point to content strings (Figure 6 B). Structural elements such as paragraphs, headings, and lists contain sequences of text nodes, and each text node can include style properties (e.g., bold, italics). We extend this framework (Figure 6C) by attaching provenance

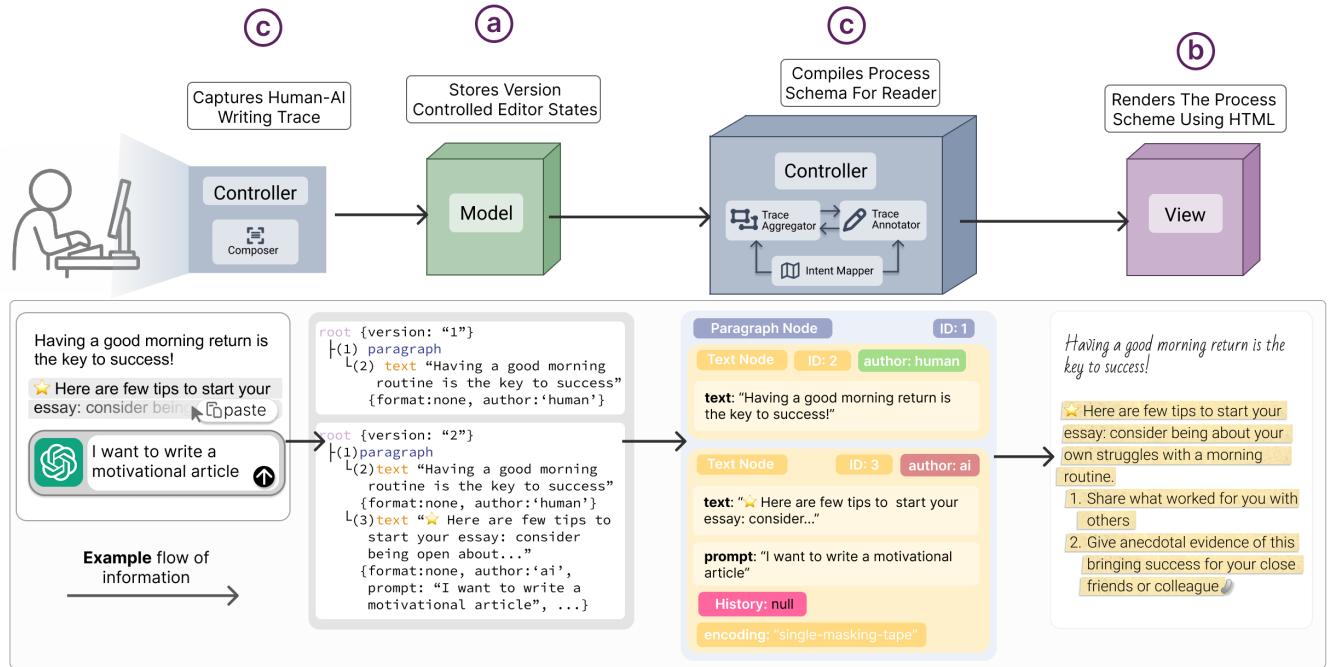


Figure 5: Overview of DraftMarks's MVC architecture. (a) The model stores the writing process, (b) the view presents the process to readers, and (c) the controller manages the bi-directional logic between model and view.

information directly to the text nodes: (1) *Author*—whether the content was written by a human or generated by AI, (2) *Prompt*—the instruction and context used when the text was AI-generated, and (3) *Generated*—the complete AI-generated text. Often writers ask AI for feedback on their content without importing that feedback into their writing canvas; these AI-generations are stored in orphan AI-authored text nodes.

Versioning in our system is event-driven rather than time-based. Whereas collaborative editors like Google Docs track synchronous and concurrent human-human edits through operation transformation [29, 69] and periodic snapshots, our model reflects the sequential nature of the human-AI workflow (Figure 6D). Current AI-assisted tools do not support concurrent human and AI editing [65, 84]. A new version is created only when: (1) an AI-authored text node is *inserted*, or (2) an AI-authored text node is *fully removed*, or (3) an AI-authored text node has *10+ characters deleted* from its content string (we use this threshold to capture micro edits). All other human revisions remain within the current version.

An example of this flow is illustrated in Figure 7. From left to right, the content evolves through the collaboration between AI and human. The corresponding editor state is shown in the next row with green and yellow highlights indicating key changes such as node insertion and version change across editor states. To ensure efficient storage, each version is represented through *references to* and not a *copy of* the structural and text nodes, enabling compact history tracking similar to Git. This model design provides a fine-grained record of human–AI collaboration, capturing both the provenance of AI contributions and the depth of the document revision. The

details regarding the capture of this knowledge representation are presented in Sect. 6. We also include the version-controlled editor representations for various types of human-AI interaction during writing in Fig. 8.

5 DraftMarks' View: Skeuomorphic Encodings

In DraftMarks, we adopt skeuomorphism as a deliberate design choice to represent human-AI writing collaboration. By incorporating visual elements into the everyday physical experience, skeuomorphic designs can increase interpretability and reduce the learning curve for new systems [43, 82]. Our technique uses several metaphors to semantically map writing behavior to visual representations. The selection of visual channels and variants for representing process information is made by the controller module (Section 6).

Masking Tape. They (Fig. 9A) indicate passages that were initially generated by AI. We selected this metaphor for AI-generated content since tape implies something provisional or temporary and externally added, qualities that align with how AI-generated content functions in co-authoring. Tape as a metaphor also allows crumpling or tearing of the tape which can visually communicate whether suggestions were kept intact, modified, or fragmented, supporting quick visual inference about the degree of human intervention. The tape also allows us to write over and add additional tape onto the new tape, all of which can signal iteration over the same content by repeated prompts. Tape specifically encodes new content generation, which was not part of the prompt passed to the AI. DraftMarks supports five different views for the masking

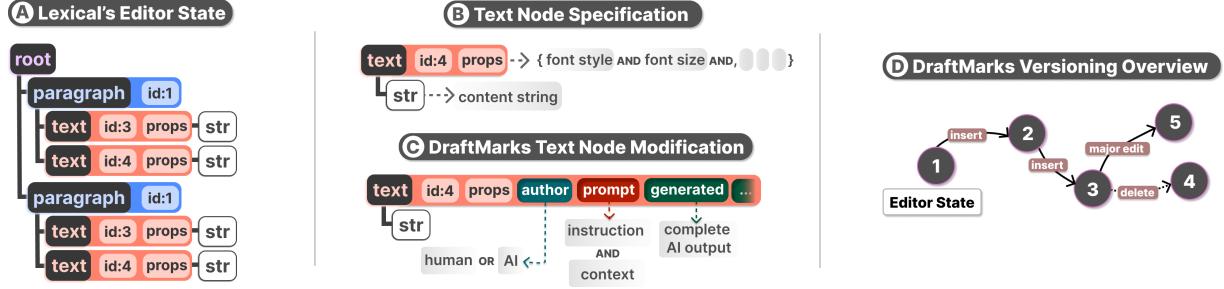


Figure 6: Here we show: (A) typical representation of editor states in rich text editors, (B) their corresponding text node, (C) text node modifications made for DraftMarks and (D) lastly overview of our version controlling architecture for capturing human-AI writing process trace.

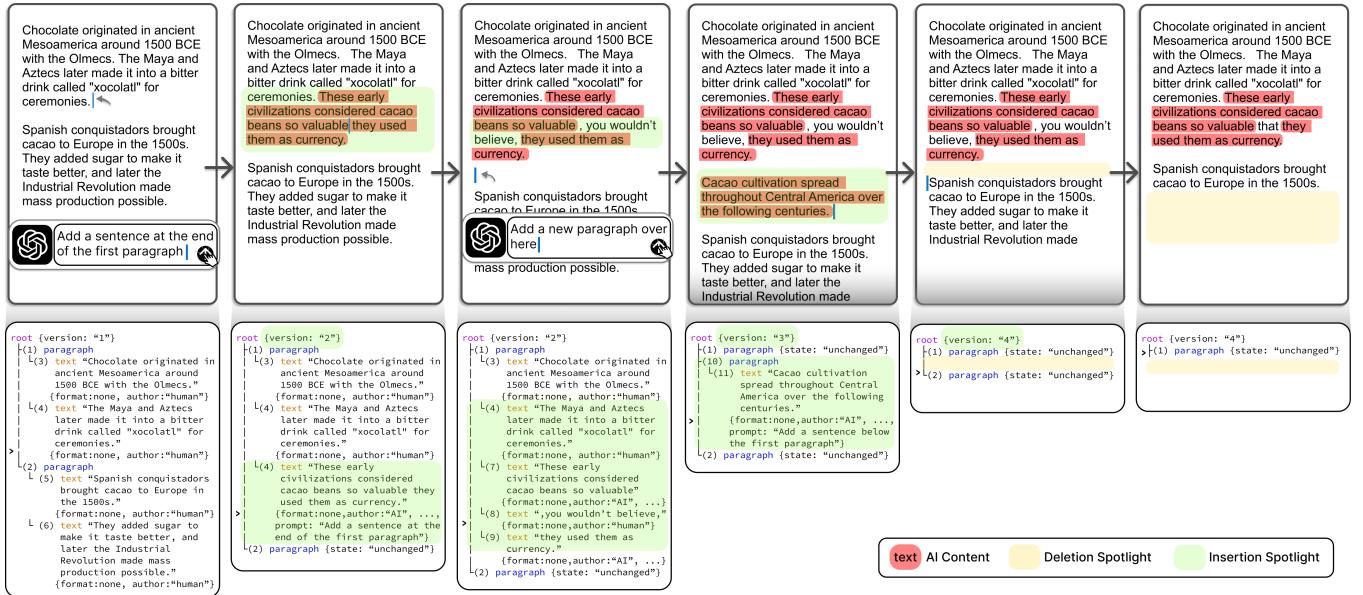


Figure 7: Here we demonstrate the changes occurring within the model of DraftMarks' on different points with human-AI writing collaboration. We ground this example in a writer using ChatGPT to write about the history of chocolate.

tape: (1) *single masking tape*: which represents a continuous strip of masking tape marking the content generated by AI, in this view granular details regarding exactly what ideas came from AI and which were already present in the prompt are not preserved as the masking tape strip marks the AI output without nuance; (2) *stacked masking tape*: represent layered tapes for iterative AI generation, trying to make AI insertions at the same spot by replacing the prior AI suggestion over and over again; (3) *scrunched masking tape*: which preserves information regarding word deletions made by humans within the AI text; (4) *torn masking tape*: which preserves information regarding word insertions made by humans within the AI text; (5) *segmented masking tape*: which preserves information regarding which phrases in the generation were AI original generation, and which ones were part of the prompt that the human wrote.

Smudge Marks. They (Fig. 9B) indicate passages that were modified by AI for tone shift, and not for the insertion of the original content. Smudges indicate areas of semantic drift, where the meaning of a sentence has changed substantially over revisions. We use smudging as a metaphor because it suggests motion, uncertainty, and transformation, qualities inherent to iterative meaning-making. DraftMarks can support two different views for the smudge mark: (1) *single smudge mark*: which represents a continuous streak of smudge over the AI's output without disambiguating precise changes made by AI; (2) *segmented smudge mark*: it preserves information regarding which phrases in the generation were AI original generation and which ones were part of the prompt that the human wrote.

Eraser Crumbs. Eraser crumbs (Fig. 9C) naturally evoke the act of erasing and rewriting by hand, activities closely related to the revision effort. In DraftMarks, the eraser crumbs appear next to AI smudge marks and masking tape to show the virtually erased

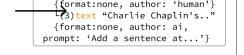
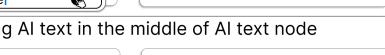
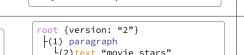
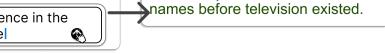
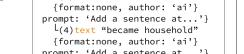
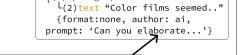
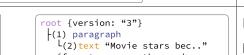
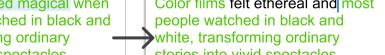
Interaction	Stored Editor States	Description
Inserting AI text at the end of a text node	 <pre>root {version: "1"} └(1) paragraph └(2)text "Color films seemed..." {format:none, author:'human'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "Color files seemed..." {format:none, author:'human'} └(3)text "Charlie Chaplin's..." {format:none, author:'ai', prompt:'Add a sentence at...'} {format:none, author:'human'}</pre>	The AI text node is appended at the end of the paragraph node, resulting in a new editor state version.
Inserting AI text in the middle of human text node	 <pre>root {version: "1"} └(1) paragraph └(2)text "Color films seemed..." {format:none, author:'human'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "Color files..." {format:none, author:'human'} └(3)text "Color films seemed..." {format:none, author:'ai', prompt:'Add a sentence at...'} {format:none, author:'human'}</pre>	At the point of insertion, the current human text node splits into two and the AI text node is inserted there, resulting in a new editor state version.
Inserting AI text in the middle of AI text node	 <pre>root {version: "1"} └(1) paragraph └(2)text "movie stars became..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "Chaplin's comedies made..." {format:none, author:'ai', prompt:'Add a sentence at...'} {format:none, author:'human'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "movie stars..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "movie stars..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(4)text "became household..." {format:none, author:'ai', prompt:'Add a sentence at...'} {format:none, author:'human'}</pre>	Similar to AI text node inside human text node the new AI text node is inserted in the split position, resulting in a new editor state version.
Replacing human text node with AI text node	 <pre>root {version: "1"} └(1) paragraph └(2)text "Color files seemed..." {format:none, author:'human'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "Color files seemed..." {format:none, author:'ai', prompt:'Can you elaborate...'} {format:none, author:'human'}</pre>	Since the human text node first has to be removed for the AI text node to replace it, that happens within the initial editor state version. A new editor state version is instantiated to capture the AI text node.
Replacing AI text node with anotherAI text node	 <pre>root {version: "1"} └(1) paragraph └(2)text "Movie stars became..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "household names before television existed." {format:none, author:'ai', prompt:'Can you elaborate this sentence?'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "Color files seemed..." {format:none, author:'ai', prompt:'Can you elaborate...'} {format:none, author:'human'}</pre>  <pre>root {version: "3"} └(1) paragraph └(2)text "Movie stars became..." {format:none, author:'ai', prompt:'Can you elaborate...'} {format:none, author:'human'}</pre>  <pre>root {version: "4"} └(1) paragraph └(2)text "Movie stars became..." {format:none, author:'ai', prompt:'Can you elaborate...'} {format:none, author:'human'}</pre>	Unlike AI text node replacing human text node, here the removal of the initial AI text is stored in a new editor state version and the insertion of the new A text node is captured within a third editor state.
Human and AI mixed iteration	 <pre>root {version: "1"} └(1) paragraph └(2)text "Color files..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "seemed magical when most people watched in black and white, transforming ordinary stories into vivid spectacles." {format:none, author:'human'}</pre>  <pre>root {version: "2"} └(1) paragraph └(2)text "Color files..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "felt ethereal and most people watched in black and white, transforming ordinary stories into vivid spectacles." {format:none, author:'human'}</pre>  <pre>root {version: "3"} └(1) paragraph └(2)text "Movie stars..." {format:none, author:'ai', prompt:'Add a sentence at...'} └(3)text "felt ethereal and most people watched in black and white, changing ordinary stories into ethereal and vivid spectacles that transported audiences to new worlds..." {format:none, author:'human'}</pre>  <pre>root {version: "4"} └(1) paragraph └(2)text "Movie stars..." {format:none, author:'ai', prompt:'Can you elaborate...'} └(3)text "people watched..." {format:none, author:'ai', prompt:'Can you elaborate...'} {format:none, author:'human'}</pre>	The insertion of human text node inside AI text node is captured within the first editor state version itself. In order to accomodate subsequent replacement of this content with new AI text node, two new version are instantiated like seen in above interactions

Figure 8: In this figure we show some common ways of integrating AI in writing. We use the example of ChatGPT being used in writing and their the corresponding overview of our version controlled data model. The text is green represent AI generation.

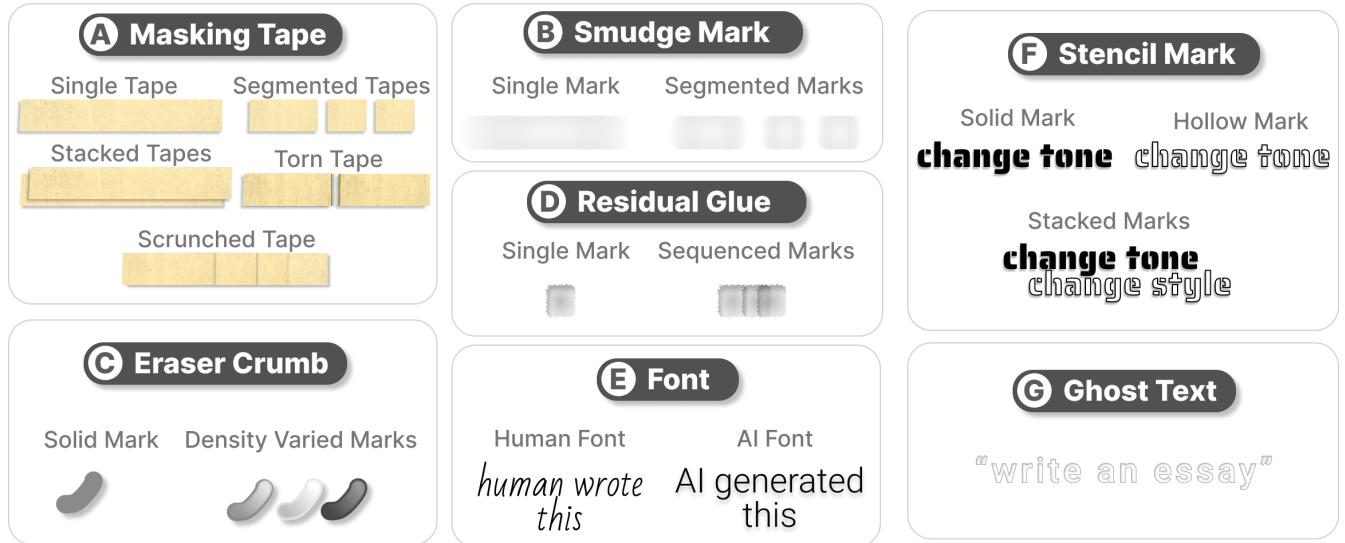


Figure 9: Marks in DraftMarks. Each addresses a unique challenge in visualizing temporal human-AI coediting.

prompts. The eraser crumb has two views: (1) *solid crumb* that has a uniform gray color; (2) *density varied crumb*: which has different shades of color. The differentiator between the two is nuance, in case of density varied crumbs the detail regarding complexity of prompts is encoded by the shade of gray.

Residual Glue. Faint, translucent glue residue (Fig. 9D) remains where masking tape was once— to indicate that an AI suggestion was initially introduced— but later removed entirely by the writer. Glue marks suggest absence with a trace, communicating that something was considered and discarded. Compared to showing the full deleted content, this subtle signal respects the integrity of the final draft while still surfacing decision points in the writing process. The views supported here are: (1) *single glue mark*: showing only the last discarded AI generation at a passage; and (2) *sequenced glue mark*: showing all versions of discarded AI generations at a passage.

Font. While not skeuomorphic, DraftMarks uses contrasting font styles (Fig. 9E) to signal the origin of the author. (1) Script or handwritten fonts for writer-authored text, and (2) sans-serif fonts for AI-generated content. Fonts provide an ambient but persistent signal that visually differentiates the origin of the text without resorting to color or other intrusive overlays.

Stencil Marks. Light, hollow, or dashed letter forms are used to mark stenciled text (Fig. 9F), passages where the writer incorporated AI feedback in a structured, guided manner. This metaphor draws from the educational practice of learning to write by tracing letters from a stencil or dotted line worksheet. This signifies that, while the structural scaffold came from the AI in the form of directed feedback, the final form was shaped by the writer. Stenciled text represents a middle ground between automation and authorship, highlighting human effort built on top of an AI scaffold. This encoding has 4 variants: (1) *single stencil mark*: a single character mark at the margin of the paragraph; (2) *layered stencil mark*: which marks multiple AI feedback generations over the same passage; (3) *dotted font strokes*: the text in the passage next to the stencil mark appears dotted and not as solid lines to depict lack of AI feedback integration done by the writer; (4) *lined font strokes*: this is the complement of dotted font strokes for cases where the writer integrates the AI feedback.

Ghost Text. Translucent ghost lines of text (Fig. 9G) show prompts or ideas that were generated, but that were not ultimately included in the final output. This metaphor suggests absence or residue - text that was once present in the writing process but was ultimately excluded or “left behind.” Ghosted text gives readers visibility into the writer’s exploratory moves without disrupting the coherence of reading the final document. Here, the two variants are: (1) showing only instructions from the prompt; (2) showing the full prompt which includes instruction as well as context.

These cues provide a visual language for DraftMarks that aligns the semantics of the writing process with familiar visual metaphors from physical writing and editing. As seen in Figure 4B—segmented smudge marks over single masking tape— the encodings can be overlayed on top of each other to represent rich breadth of complex writing and editing workflows. Different combinations enable the human-AI writing process to be traced at different levels of depth, depending on the importance given to the final content, behind-the-scenes prompts, iteration, or precise phrase-level provenance.

Each of these metaphors was selected through experimentation via encoding data and deliberation amongst the authors, both for its recognizability and its ability to encode the corresponding process.

6 Controller Logic: Generating Skeuomorphic Encodings from Data Model

Reading practices are fundamentally *contextual*. Different stakeholders have different interpretive frameworks and evaluative priorities. Teachers engage in pedagogically-oriented reading — simultaneously assessing content comprehension and student learning processes — while academic reviewers focus on novelty, methodological rigor, and scholarly contribution. These stakeholder-specific requirements should be taken into account when visualizing human-AI writing collaboration. In other words, it is difficult to formulate a singular augmentation algorithm that can completely capture all insights different readers might need while also remaining legible to the reader. Therefore, in DraftMarks we take a human-centered approach to design and develop the various algorithms for mapping the data model to role-dependent views. These algorithms are implemented in DraftMarks’ controller module (Figure 5c).

6.1 Formative Study for Designing Controller Logic

To better understand reader needs, we conducted a formative study with three different stakeholders: (A) teachers, (B) academic reviewers, and (C) general readers. From this diverse pool of readers ($N = 21$, i.e., 3 stakeholder \times 7 participants), we sought to understand what visibility into the writing process they might need, and the associated comprehension and assessment-related decisions they would make. To conduct the study, we developed a baseline version of DraftMarks implementing the different skeuomorphic augmentations mentioned in Section 5. This version of the system is intended as a design probe for the study. DraftMarks implemented *masking tape* to represent AI content, *ghost text* to represent prompts, and additionally *eraser crumb*, *smudge marks*, and *residual glue*, along with different fonts for human-written and AI-generated content. This baseline tool reflects a deliberately simple implementation, guided by heuristics and our design judgment, to surface possibilities rather than optimize performance.

Participants. We recruited participants via LinkedIn, Twitter, and word of mouth. For *teachers*, we had the selection criteria of being a teacher currently and having long-form writing (narrative or argumentative) assessments as part of their curriculum. All teachers recruited were from private schools with relatively small class sizes (20 classrooms of students). For *reviewers*, we had the selection criteria of being a senior Ph.D. student who has reviewed at least three conference papers. For *general reader*, we recruited participants who did not meet the inclusion criteria of the other two stakeholders. All details regarding the participants can be found in Table 1. Each participant was compensated with \$30 for their time. The institution’s IRB approved the study.

Study procedure. Each study session lasted no more than one hour. At the start of each session, participants received a live demonstration of DraftMarks on an example text that showcased all the basic skeuomorphic encodings and their associated interactions. After

(A) Teachers				
ID	Experience	Discipline	Grade Level	AI Acceptance
P1	10+ years	History	Middle School	3
P2	10+ years	English	Middle & High School	5
P3	7 years	English	High School	5
P4	10+ years	English	Middle School	3
P5	7 years	English	Elementary School	4
P6	9 years	History	Middle & High School	5
P7	10+ years	English & Math	Middle School	4
(B) Academic Reviewers				
ID	Experience	PhD Year	Discipline	AI Acceptance
R1	3 papers	3rd Year	HCI	4
R2	10+ papers	6th Year	Ubicomp	3
R3	10+ papers	3rd Year	Robotics	2
R4	6 papers	3rd Year	Machine Learning	4
R5	3 papers	3rd Year	Machine Learning	3
R6	7 papers	5th Year	HCI	4
R7	6 papers	4th Year	HCI	5
(C) General Readers				
ID	Reading	Education	Profession	AI Acceptance
S1	Very often	High School Diploma	Undergraduate Student	4
S2	Very often	High School Diploma	Undergraduate Student	4
S3	Often	High School Diploma	Undergraduate Student	5
S4	Somewhat often	Bachelor's Degree	Engineer	4
S5	Very often	High School Diploma	Undergraduate Student	5
S6	Very often	Master's Degree	PhD Student	5
S7	Occasionally	High School Diploma	Undergraduate Student	5

Table 1: Participant demographics from our controller design probe (Sect. 6.2.2). “AI Acceptance” refers to responses on whether AI use in writing is acceptable, rated on a 5-point scale (1 = strongly disagree, 5 = strongly agree).

this introduction, participants were asked to assess a piece of writing—co-written with AI—relevant to their readership. The details regarding essay selection are provided in the following paragraph. The participants were asked to think aloud about their assessment and comprehension. After this activity, participants received a general rubric to assess the writing on various dimensions. The purpose of the rubric scoring was only to understand their evaluative decision-making processes. Participants were then shown the video recording of the actual writing process from the authors of the essays to elicit any additional information they would have liked to see in DraftMarks to help their comprehension and assessment needs. Lastly, we administered a demographic survey and a short questionnaire specific to each readership.

Essay Selection. For the study task, we curated role-specific essays. For teachers, we selected a writing session from the CoAuthor dataset [42], which captures interactions with a tool that generates four alternative sentences as continuations to the current text based on writer-initiated triggers. For reviewers, we recruited a postdoctoral fellow who wrote an article aimed at a trade magazine audience—in this case, ACM Interactions—based on their dissertation. The participant used Script&Shift [65], an AI co-writing tool that employs a collage-based metaphor with multiple “AI-helpers” specialized for different writing subprocesses. After a demonstration of

Script&Shift, the participant completed the task and received a \$30 honorarium. For general readers, we recruited a master’s student to write an article on “privacy in the digital age” with the assistance of ChatGPT. In the case of CoAuthor, we obtained the writing session video from the project’s website. For Script&Shift and ChatGPT, we recorded participants’ screens while they wrote. Additionally, for the recruited writers, we ran the DraftMarks composer in the background of their browser to capture the data model.

Data Analysis. We conducted reflexive thematic analysis [11]. Two of the authors collaboratively coded six transcripts and video recordings to develop an initial codebook, which was then applied to the remaining transcripts. At this phase, the coding was done independently by the authors using the initial code book. The authors were invited to come up with new codes that were not observed in the initial subset. The final codebook, devised by the authors, was combined into a single, cohesive one after a round of discussion (number of top-level codes = 19). An independent rater not affiliated with the article then assisted in the process of combining, splitting, or removing existing codes (number of top-level codes = 16). The final code book constituted the evaluative and interpretative process traces we were looking for, as shown in Figure 10.

Following this activity, the two authors and the independent rater mapped these codes to the different variants across the visual encoding channels until agreement was reached. Three codes: (QJ) quality judgment, (AL) applicability lens, and (EC) expectation calibration were excluded from this process. QJ was excluded because it was directly related to the task and, therefore, was evenly present in high frequency for all participants. We exclude AL as this was the participants who commented on how DraftMarks could be helpful for different applications, not necessarily offering any visual encoding or specific insight into human-AI collaboration. Lastly, we exclude EC because of their reaction to seeing the video. All remarks about DraftMarks and their reading needs during the video showcase were captured in different codes.

6.1.1 Study Results. Overall, we observed that teachers were more curious about the trace behind the written content than the other two stakeholders. Reviewers shared that they often have limited time to review papers, and the trade-off between cognitive load and insights was not justifiable for them in most instances. They did say that provenance of ideas is helpful and expressed desire for more summative insights in contrast to formative pieces in DraftMarks. Lastly, for general readers DraftMarks had an impact on their trust and willingness to read to article, sharing they value authenticity and originality in author's voice and DraftMarks' traces help capture that.

Insights on Teachers. We observed significant frustration among teachers regarding the monitoring and guidance of student AI use in assessments. P6 described existing detection mechanisms as unhelpful, noting that students use so-called "humanizers" to circumvent tools like GPTZero. However, teachers demonstrated pragmatic acceptance of AI as an educational reality. P6 shared, "*I came to the conclusion that [AI] is here to stay, that it is powerful. I am using it! [...] If I open the conversation up and let them show me how they're using it, I can start to figure out how to help them.*" Teachers emphasized that AI should support rather than replace student thinking. P4 explained, "*The point of writing is to express your thinking—to make visible the invisible thinking. So AI should help you make your thinking visible. It shouldn't do your thinking for you.*" With respect to DraftMarks, teachers valued its formative assessment potential. P2 shared that the tool "*makes it clear what the student is struggling with from a formative point of view. That to me feels like a point that I like as a teacher.*" However, teachers noted that such detailed process analysis requires manageable class sizes and sufficient grading time. Teachers' emphasis on formative assessment and process reconstruction (PR in) directly informed our controller design. Since teachers frequently engaged in pedagogical lens evaluation (PDL in Figure 10) and connected AI traces to quality judgments (PQL in Figure 10), we implemented multiple encoding variants for each channel to support granular process analysis (Figure 11). For example, teachers' interest in students' decision-making around rejected AI suggestions (residual glue) led us to preserve these traces for even when content was discarded. The controller's aggregation depth for teachers is set to capture fine-grained interactions because, as P2 noted, understanding "*what the student is struggling with*" requires visibility into micro-level writing decisions. The teachers shared they had been experimenting with revision surfacing tools [12, 59] for that micro-level decision making visibility.

Insights on Reviewers. Academic reviewers expressed considerably less concern about AI usage in manuscript preparation, consistently emphasizing that evaluation should focus on the final writing product rather than the writing process. R7 articulated this stance clearly: "*I'm here to review a piece of writing. [...] this seems like extra work.*" Reviewers demonstrated pragmatic acceptance of AI usage, with R7 stating they weren't bothered by generative AI in research papers. However, they distinguished between different types of assistance, explaining: "*I don't care about AI in the writing, I care about AI in the idea phase.*" This reflects reviewers' concern with intellectual contribution over mechanical writing assistance. Despite general resistance to process-focused tools, some reviewers recognized potential value in specific circumstances. R6 noted it "*would be helpful to have differentiation between when AI produces new ideas and when it is just helping polish content,*" particularly for non-native English speakers using AI for grammatical corrections. Reviewers' focus on intellectual contribution over process led to a minimal controller implementation. Their primary concern with author attribution of ideas (AA) drove our decision to highlight only content generation that might represent new intellectual contributions versus mechanical assistance. The sparse mapping in Figure 11 reflects reviewers' explicit preference against process exploration (ME in Figure 10)—we preserve only traces that support their core review obligations. For instance, iteration depth markers help reviewers quickly identify content the author invested significant effort in (PR), while other process details are suppressed to avoid the "extra work" R7 described. The controller prioritizes efficiency, surfacing only process cues that influence reviewer judgment (MI) without requiring deep engagement with collaborative traces.

Considering the overall mild interest in process trace of human-ai co-writing, our mapping of the reviewer's intents to visual encoding variants is quite sparse (Figure 11). For support their review process, they are concerned with author attribution of idea (AA in Figure 10). They care about process reconstruction (PR) to see if there phrases/sentence the author paid special interest to (high iteration depth). They seemed content relying on process cues to influence their judgment (MI in Figure 10) but largely believed any further process exploration through marks would be unhelpful (ME in Figure 10) as it takes away from their reviewing time.

Insights on General readers. For general readers, AI acceptance was contingent on transparency and authorial effort. Unlike professionals with specific obligations, readers expressed concerns about authenticity and personal investment. S6 articulated this sentiment: "*I don't mind AI tools [...] But it should be used as supplement to your writing, and not... It shouldn't do the writing for you.*" Readers valued human voice and perspective, with S6 explaining they read articles "*to listen to people*" but extensive AI generation made them feel they weren't "*listening to someone's point of view*" but rather "*what a chatbot wrote back to them.*" The perception of effort significantly influenced reader trust and engagement, with low-effort AI use raising questions about whether the content was worth reading. Regarding DraftMarks, readers found the visualization valuable for understanding the writing process and desired comprehensive information about AI interactions, including access to original prompts and AI responses. General readers' authenticity concerns and trust responses were key in designing their controller. The dramatic

CID	Code Name	Code Description	Code Example	Code Frequency
AA	Author Attribution	Who wrote what (human/AI/uncertain)	"So it seems like this was mostly written by AI" [R2]	4.4 6.3 3.6
IE	Effort Inference	Judgement about work/effort invested	"That again shows me that the student [...] isn't blindly just taking the AI suggestion [...] but like really caring about their work" [P3]	2.1 3.6 4.4
PR	Process Reconstruction	Attempts to understand how text was created	"It's almost like they were trying to use AI to connect each thought in their thought process, for the story." [P7]	8.8 5 4.4
QJ	Quality Judgement	Evaluative statement about text quality	"I would never have someone write a essay that starts with [...]" [P1]	10.4 8 8.3
PQL	Process Quality Link	Explicitly connecting process information to quality assessment	"when [...] it is simply enhancing what the human already wrote [...] I think that the AI actually improved [...] it much clearer" [R5]	6.3 2.8 2.1
ME	Mark Exploration	Active investigation of marks (clicking/hovering)	"So I can click the residual glue to see the original AI suggestion? That sentence seems kind of out there." [P5]	3.1 4.7 4.6
MI	Mark Influence	Encoding explicitly changes interpretation or behavior	"This student used AI wisely to enhance their response, not to do the writing for them." [P6]	3 4.4 4.8
MO	Mark Overwhelm	Too much visual information or confusion	"I think what's hard here is things are nesting inside of other things, and so I'm getting lost in what writing is the writing" [R6]	0.8 2.1 0
CS	Credibility Shift	Trust changes based on process information	"However, I feel if I were to have higher trust, it should be fully written [by human], or just solely using AI for syntax?" [S1]	0.6 3.1 4
AC	Authenticity Concern	Questions about originality or voice	"I'm trying to get them to not feel like this is a secret we have to hide stuff. I want them to be honest about their process" [P1]	4.4 0 3.3
CB	Coherence Building	Connecting ideas within text or process flow	"I thought the author was more relaxed in writing, but now the AI here is not consistent in the way of writing" [R2]	0.2 1 0
IS	Information Seeking	Looking for additional context or information	"But that paragraph's evolution isn't anywhere here, so this is from a previous draft that I can't see?" [R6]	6.3 1.4 3.1
CL	Contribution Lens	Analyzing originality and scholarly merit of content	"I care about AI in the, idea phase [...] I would question using AI to do all of the [...] intellectual lifting." [R1]	0 1 0.6
PL	Pedagogical Lens	Analyzing teaching or learning concerns	"I see they used AI for some dialogue, so they need help [...] creating dialogue for characters to get the next point in the story." [P5]	6.4 0.6 0.6
AL	Applicability Lens	Analyzing usefulness or applicability of tool	"[DraftMarks] will be very useful for reflection or to compare [LLM use in] my writings" [R1]	0.4 1.4 0.9
EC	Expectation Calibration	Surprised by or confirming assumption upon seeing video recording	"I don't think [DraftMarks] is misleading. Seeing the edits [...] it provides a bigger picture of how [AI] is used" [S5]	0.4 2.1 2.7

Figure 10: This figure showcases the final code book devised from thematic analysis of all 21 transcripts across stakeholders. The code frequency column contains average frequency of the code across participants for teachers (red), reviewers (green), and general readers (blue).

trust reduction we observed when readers saw masking tape (as S6 demonstrated) led us to provide both smudge marks and masking tape encodings to them—distinguishing between tonal refinement and novel content generation (Figure 11). The aggregator maintains higher temporal depth for general readers than reviewers, reflecting their willingness to explore process details when transparency supports their authenticity assessments.

Writer Perspective. Both reviewers and readers saw DraftMarks as a powerful metacognitive tool for writers themselves. Since writers often struggle to gauge AI effectiveness, process visualization becomes crucial for improving future collaboration. R1 explained: “Well, it might be very useful for the authors [...] I can even do the reflection and compare my writings to what the LLMs added [afterward].” The tool also addresses temporal challenges in writing workflows. S5 observed: “if I write this, and then I come back, like, a week or two weeks later, I’m not sure I’ll remember what I said [in prompts], right?” However, stakeholders emphasized that writers should maintain control over captured content, noting privacy concerns about exposing all writing processes.

6.2 Controller Implementation

The controller facilitates bi-directional communication between the model and the view as shown in Figure 5. The controller is responsible for performing mutations to the underlying data structure of DraftMarks based on the input of the user (Section 6.2.1).

This is the part of the system where our version-controlled editor states are manipulated, corresponding to human-AI collaborative writing activity. The controller is also responsible for translating this data structure into a format ready to be rendered by the view and interacted with by the user (Section 6.2.2).

6.2.1 View to Model. The capture of the data structure is handled by the composer module. The composer in DraftMarks is an extension of Lexical’s editor state listener. The critical implementation artifact from DraftMarks is the inclusion of a versioned history manager and content provenance (human or AI) tracker. This module can store writing trace between AI and human for several common AI co-writing setups: (1) split context (Google Doc with ChatGPT), (2) integrated co-writing tools (Script&Shift, ABScribe [60]), and (3) ambient AI assistants (Grammarly chrome extension). In line with findings around writer privacy in our controller formative study, the composer requires explicit user permission for writing tracking, preventing tracking of content they are not comfortable sharing. The composer tracks keystroke-level data; content it can map to keystrokes is marked as human content, whereas that which it cannot is considered AI-generated. Copy and pasted content that is not local to an app is also considered AI-generated. Through these heuristics, the composer acquires the ground-truth data model, which the model-to-view controller then processes to determine final visual artifacts.

	Teacher			Reviewer			General		
	Encoding(s)	Intents	Encoding(s)	Intents	Encoding(s)	Intents			
Masking Tape	Scrunched	PL PQL IS AC AA ME MI	Single Stacked	AA PR ME EI IS CB CL	Single	MI CS AA AC			
	Torn	AC AA ME							
	Segmented	MI							
Smudge Mark	Segmented	PL PQL AC AA MI			Single	MI EI CS AA AC			
Font	Script	PL AC AA MI EI	Script	AA MI EI CS CB CL	San Serif				
	San Serif		San Serif						
Eraser Crumb	Density Varied	PR PQL IS ME MI EI	Density Varied	PR ME MI EI CS	Solid	MI ME PR EI CS AC			
	Solid	PL PQL IS							
	Hollow	ME MI							
Stencil Mark	Stacked								
	Instruction Context	PR PL PQL IS AC AA ME MI EI	Instruction Context	PR MI EI CS	Instruction Context	MI PR EI CS AC IS			
Residual Glue	Single Sequenced	PR PL PQL IS AC AA							

Figure 11: This figure encapsulates findings from the formative assessment. For each stakeholder, the visual encoding variants available to render are listed along with the corresponding reading intent they satisfy. This logic drives the intent mapper in the Model-to-View Controller

6.2.2 Model to View. Based on the findings from our formative study, we devised context-specific algorithms for converting the data model into view ready representations. Using our codebook, we distilled a set of priorities for each stakeholder and based on them identified mapping to encoding variants within the various visual encoding channels of DraftMarks. Our mapping between encoding channels and stakeholder is shown in Figure 11.

The controller transforms version-controlled editor states into a Process Schema—a monolithic nested structure that encapsulates all relevant traces for the reading stakeholder. This transformation is orchestrated by three interconnected components: the Trace Aggregator, Trace Annotator, and Intent Mapper, which work together to create a stakeholder-specific representation optimized for visualization. An example flow of decision making within the controller for different reader types is shown in Figure 13

Intent Mapper. The *Intent Mapper* serves as the configuration hub, defining transformation parameters based on stakeholder profiles identified in our formative study. Teachers need granular access to student-AI interactions to assess learning and provide feedback, while general readers prefer streamlined views that highlight key collaborative patterns without overwhelming detail. The Intent Mapper maintains mappings between stakeholder intents and available visual encoding channels and their variants, communicating depth requirements to the Trace Aggregator and encoding permissions to the Trace Annotator. Details regarding this are present in Figure 11. As shown in row 1 of Figure 12, in case of reviewers the AI iteration is represented with stacked masking tape. The same

iteration is rendered differently for teachers and general reader where further subdivision between tonal-shift and new-content generation type AI collaboration can be differentiated. In case of row 2 in Figure 12, while reviewers and general reader have the same encodings, teacher is able to capture the discarded suggestions using residual glue.

Trace Aggregator. The *Trace Aggregator* determines which details from the version-controlled editor states should be preserved in the Process Schema (see Figure 13). Rather than storing complete version histories, it makes selective decisions about temporal depth (how many versions back to preserve traces), granularity level (which text nodes warrant individual tracking), and nesting structure (how to organize traces hierarchically). The significance of changes and stakeholder type drive these decisions. At each inclusion/exclusion decision point, the aggregator consults the *Intent Mapper* to ensure the preserved detail level aligns with the stakeholder requirements, then invokes the Trace Annotator (yellow diamond shape in Figure 13) to determine the appropriate labeling.

Trace Annotator. The *Trace Annotator* analyzes AI-authored text nodes to classify the nature of human-AI interactions and assigns appropriate visual encodings. It distinguishes between iterative and novel generation requests by examining prompt content and AI-generated text. Iterative AI calls that involve sequential requests for similar content can be represented as stacked masking tapes. Novel requests involving different content types receive separate visual elements. The annotator also identifies different generation

types: new content generation is encoded with masking tape, tonal refinement uses smudge encoding, removed AI content appears as residual glue, and used prompts become clickable eraser crumbs that reveal ghost text. The annotator also passes to the Aggregator details about the consolidated structure of some content. Throughout this process, the annotator consults the Intent Mapper to ensure selected encodings align with the stakeholder's comprehension needs.

7 Evaluation

Our study sought to evaluate the effectiveness of DraftMarks' visual encodings in surfacing human-AI writing collaboration patterns and enhancing users' comprehension and assessment abilities when reviewing AI-assisted content. We used a between-subjects design to compare our augmented reading tool against a baseline interface (described in Section 7.2)

7.1 Participants

We conducted the study through Prolific and screened to select participants who regularly engage in written content professionally, with roles including teacher, journalist, and copywriter. We also required participants to have a 95% or higher approval rating on Prolific to ensure high-quality participation. We recruited 70 participants ($F=36$, $M=32$, $NB=2$), with 35 participants randomly assigned to each condition. Participants reported ages between 18 and 60 years and older, with the largest group (20 participants) in the age range of 31-40. Regarding AI usage in their own writing, most participants reported using AI sometimes (25) or often/very often (31), with only 2 participants reporting never using AI tools.

7.2 Study Procedure

Each study session lasted approximately 30 minutes and consisted of three phases: interface tutorial, practice task, and main evaluation task. The main task required participants to comprehend a 600-word essay about "Need for Social Media Reform for Children's Use" that was written collaboratively with ChatGPT and exhibited extensive human-AI interaction patterns. The participants then answered comprehension questions related to the content of the essay and the collaboration patterns. The control condition used a split view interface that shows the essay with AI-generated sentences highlighted alongside the available chat log. The treatment condition used DraftMarks with enhanced visual encodings for human-AI collaboration patterns. Following the main task, the treatment participants completed the Cognitive load survey and the system usability scale questionnaire. Participants in both conditions completed the demographic survey, filled out 3 transparency questions on a 7-point likert scale: (Q1) *"I feel confident in my assessment of the essay"*, (Q2) *"I felt I had enough information to assess the author's writing process"* and (Q3) *"I was able to perceive the human effort in this essay"* and provided open-ended feedback to these questions: (F1) *"What, if anything, helped you most in understanding the human–AI collaboration?"*, and (F2) *"What, if anything, was confusing or misleading?"*. Each participant received \$6 compensation for their participation.

7.3 Results

7.3.1 Comprehension Questions. We analyzed question response accuracy using Fisher's Exact Test for individual questions (due to below 5 frequencies for correctness/incorrectness in some questions) and the Mann-Whitney U test for aggregate performance across both conditions ($N = 35$ per condition). The treatment condition significantly outperformed the control in three individual questions. Question 1 improved dramatically from 2.9% to 51.4% correct ($p < 0.001$), Question 3 from 11.4% to 48.6% ($p < 0.001$), and Question 2 from 17.1% to 40.0% ($p = 0.026$). The pattern of improvements suggests that DraftMarks is particularly effective at helping users track specific editorial changes and AI contributions. Question 1's dramatic improvement (from 2.9% to 51.4%) indicates that our visualization successfully addresses a fundamental challenge in understanding AI-assisted writing: identifying where changes occurred. The substantial gains in Questions 2 and 3, which focus on understanding AI feedback integration, suggest that making the revision process visible helps users develop better mental models of human-AI collaboration. However, the relatively low baseline performance across all questions (2.9%–17.1%) confirms that without appropriate tools, readers struggle significantly to understand AI's role in collaborative writing, making transparency tools like DraftMarks essential rather than merely helpful. In general, participants in the treatment group scored significantly higher than controls ($\mu = 4.29$ vs. $\mu = 2.86$ out of 7, $U = 335.5$, $p < 0.001$, $r = 0.52$), representing a 50% improvement with a large effect size.

7.3.2 Cognitive Load Survey. Based on our analysis of the CLS responses, participants reported moderate intrinsic load ($\mu = 4.31$, $\sigma = 2.66$) and low extraneous load ($\mu = 3.03$, $\sigma = 2.46$) while using the system. Furthermore, their rated self-perceived learning, captured by germane load, was high ($\mu = 7.42$, $\sigma = 1.96$). This cognitive load profile suggests that DraftMarks successfully balances information richness with cognitive accessibility. The moderate intrinsic load indicates that understanding AI-assisted writing processes requires meaningful mental effort, which is appropriate given the complexity of human-AI collaboration. The high germane load with relatively low variability ($\sigma = 1.96$) suggests consistent learning experiences across participants, indicating that the system effectively supports comprehension regardless of individual differences in technical background or reading strategies.

7.3.3 System Usability Scale. DraftMarks achieved a strong average SUS score of 80.5 ($\sigma = 12.8$), placing it well above the 68 threshold for acceptable usability. Analysis of individual SUS items reveals that users found the system easy to use (SUS 3: 74% agreement) and expressed confidence in using it (SUS 9: 77% agreement). However, some participants indicated they would need technical support to use the system (SUS 4: 69% disagreement with needing support) and found various functions well integrated (SUS 5: 89% agreement). The strong SUS score is particularly meaningful given that participants used DraftMarks with no prior training, suggesting the visualization design successfully leverages familiar reading and annotation paradigms. The high agreement on function integration (SUS 5: 89%) indicates that our approach of embedding transparency features directly into the document interface feels

Human-AI Collaborative Editing	Reviewers	Teachers	General Reader
Replacing AI text node with anotherAI text node Movie stars became household names before television existed. Can you elaborate this sentence?	Movie stars became household names across America before television brought entertainment into homes.	Movie stars became household names across America before television brought entertainment into homes.	Movie stars became household names across America before television brought entertainment into homes.
Human and AI mixed iteration Color films seemed magical when most people watched in black and white, transforming ordinary stories into vivid spectacles. Color films felt ethereal and when most people watched in black and white, changing ordinary stories into ethereal and vivid spectacles that transported audiences to new worlds... Can you elaborate this sentence?	people watched in black and white, changing ordinary stories into ethereal and vivid spectacles that transported audiences to new worlds ...	people watched in black and white, changing ordinary stories into ethereal and vivid spectacles that transported audiences to new worlds ...	people watched in black and white, changing ordinary stories into ethereal and vivid spectacles that transported audiences to new worlds ...

Figure 12: In this figure we show the more complex interactions from Figure 8 (bottom two) and show what the view for each stakeholder would look like determined by the controller.

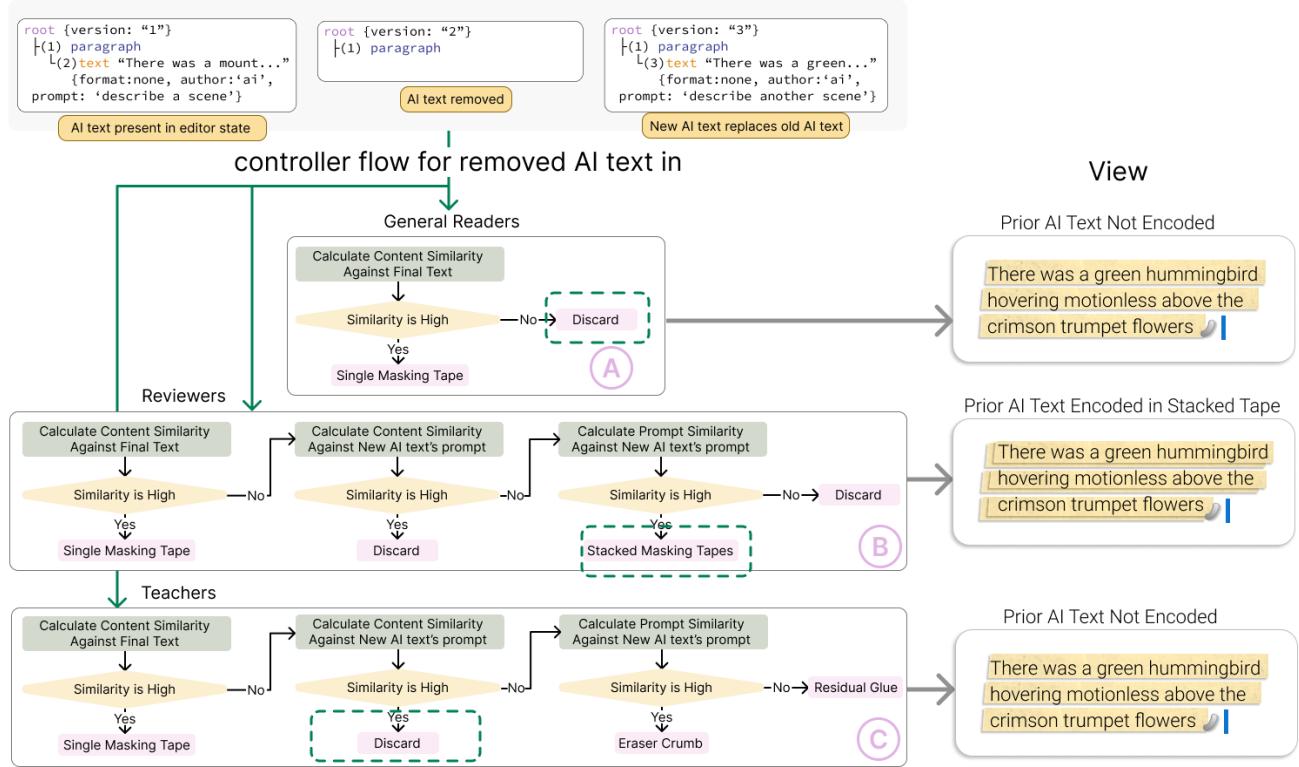


Figure 13: In this figure we show the controller's decision making trace from the model to the view. We illustrate how a previously discarded text node is aggregated and annotated for rendering based on the readership. The dashed green rectangle represented the rendering decision made for each reader type.

natural to users. However, the mixed responses on technical support needs (SUS 4) suggest that while most users can operate the system independently, the complexity of tracking AI contributions may initially overwhelm some users.

7.3.4 Transparency Measure. Analysing the three transparency questions, we notice participants in both conditions reported relatively high transparency perceptions with the DraftMarks condition consistently outperformed AI-Highlight marginally across all. For *confidence in assessment*, DraftMarks participants scored higher ($\mu = 5.81$, $\sigma = 1.00$) compared to AI-Highlight ($\mu = 5.27$, $\sigma = 1.26$).

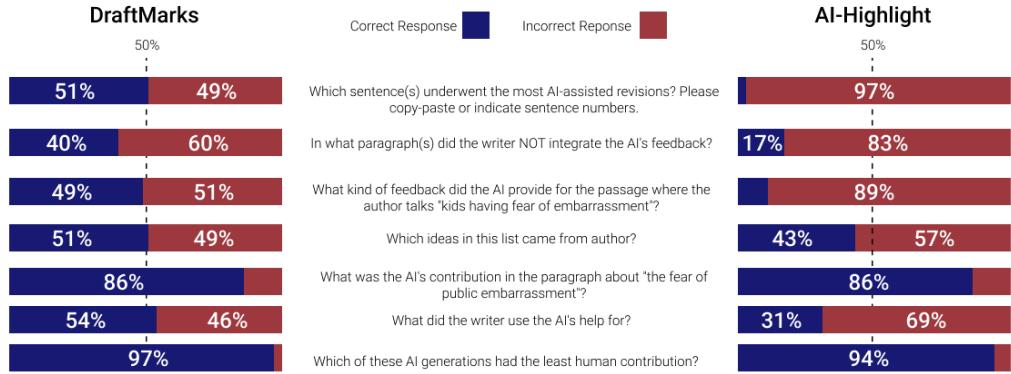


Figure 14: This figure showcases the distribution of correctness for the comprehension questionnaire between AI-Highlighted and DraftMarks condition

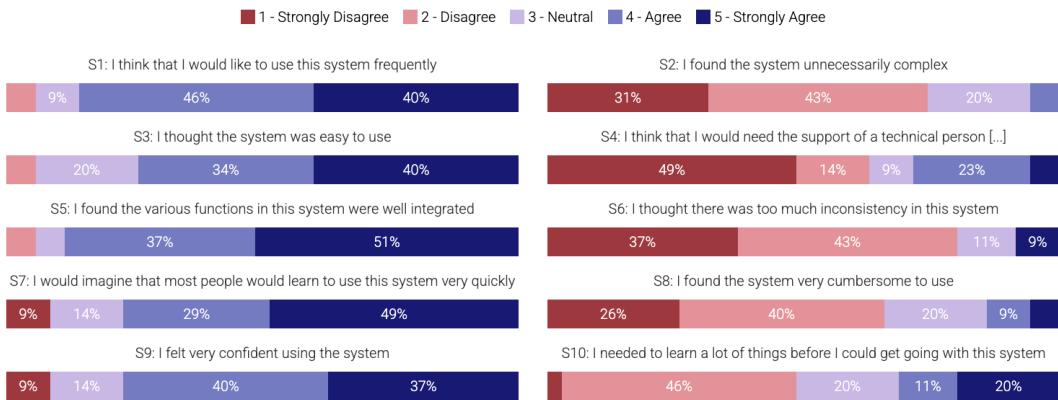


Figure 15: This figure showcases the distribution of SUS ratings (1-5) across the 35 participants for DraftMarks. For items in the left column, higher is better, and for items in the right column, lower is better.

Similarly, for *information sufficiency*, DraftMarks again showed superior performance ($\mu = 5.74$, $\sigma = 1.32$) versus AI-Highlight ($\mu = 5.24$, $\sigma = 1.09$). The *perception of human effort visibility* followed the same pattern, with DraftMarks ($\mu = 5.93$, $\sigma = 1.04$) outperforming AI-Highlight ($\mu = 5.48$, $\sigma = 1.18$). Despite these modest differences in transparency perceptions, participants in the AI-Highlight condition performed significantly worse on comprehension measures, suggesting that surface-level visibility of AI assistance (through chat logs and highlighted text) created an illusion of process understanding without meaningful insight into the writing workflow itself.

7.4 Qualitative Feedback

The open-ended responses revealed distinct patterns between conditions regarding what helped users understand human-AI collaboration. For DraftMarks condition 74% (26/35) participants shared their thoughts for F1 and F2, while for AI-Highlighted 91% (31/35) responded. For F1, over half of DraftMarks respondents (15/26, 57%) praised eraser crumbs, residual glue and masking tape for helping them see what stuck in the writing from AI and what didn't. In

contrast, 31% of AI-Highlight participants (10/32) experienced confusion about attribution, with one participant noting “[It was] hard to tell how much help the AI gave or where the human made changes”. While only 8% of system users (2/26) experiencing confusing, one respondent requested clearer legends for the process cues. The 35% of AI-Highlight users who struggled did so with fundamental questions of “figuring out exactly which parts were written by the AI and which by the human.”

8 Discussion and Future Work

Based on our design and evaluation, here we reflect on the implications of DraftMarks for readers and writers. Based on our reflection, we identify tensions and opportunities that warrant further investigation.

8.1 The Cost of Making the Writing Process Legible

Writers have long used signaling as a *deliberate* rhetorical strategy to shape the way their work is read [46]. Titles, headings, and typographic emphasis such as **bold** and *italicized* text help readers infer importance, intent, and argumentative flow [46]. These signals are authored under the writer’s control and serve their communication goals. In contrast, DraftMarks introduces a form of implicit signaling: cues that are automatically generated from interaction data and surfaced within the text to make the writing process legible. The result is a document that not only communicates ideas but also reveals how those ideas were constructed, who revised, who delegated, and how much cognitive effort each section reflects. This form of visibility offers clear benefits. It supports more transparent forms of authorship, helps readers assess originality and deliberation, and enables richer feedback conversations between writers, instructors, collaborators, or reviewers.

However, there are other trade-offs to consider when making the process implicitly visible. Because these signals are not author-curated, they may surface aspects of writing that feel personal, incomplete, or even compromising, such as moments of hesitation, minimal revision, or heavy reliance on AI-generated content. With DraftMarks, what was once invisible now becomes part of the public artifact. In an educational and evaluative setting, this visibility can redefine the way effort, originality, and ability are perceived. The core trade-off is between *interpretive transparency* and *authorial agency*, i.e., helping readers understand how text was constructed and giving writers control over what is disclosed. This raises important design questions for DraftMarks that need to be addressed in future work. Should writers be able to redact or annotate process signals? Can visibility be audience-specific, e.g., shown to instructors but hidden in publication? And how might authors be supported in interpreting and managing their own signals before others? Questions about who decides what becomes visible, to whom, and when, are important topics for future exploration.

8.2 Supporting Writer Reflection

Although DraftMarks is designed to help readers interpret AI-assisted writing, it has the potential to offer direct benefits to writers themselves. In traditional pen-and-paper writing, the act of drafting leaves behind a trail that helps writers reflect on their processes and decisions. In contrast, AI-generated text is often presented to writers fully formed and blended within the context of the current text, making it easy to accept suggestions without revisiting or refining them. This can potentially erode opportunities for deliberate thinking, reflection, and self-assessment. DraftMarks can alleviate this concern by externalizing writers’ interactions with AI and revealing where their effort has (or has not) been directed. Visual cues such as eraser marks, prompt iteration indicators, and masking tape overlays make the invisible parts of their process tangible. This visibility can encourage writers to pause and reflect on questions such as *Have I done enough to make this idea my own? Am I critically engaged with AI suggestions, or simply accepting them?* By surfacing patterns of revision, delegation, and iteration, DraftMarks helps writers monitor their writing habits and prompts

more intentional engagement with the text. This type of metacognitive support is especially beneficial in educational settings. For students, DraftMarks can promote awareness of their development as writers, providing a concrete record of revision that supports growth and feedback conversations (e.g., [31]). Future work can explore specific interactive affordances and guidance for writers when co-authoring with AI.

Our approach also shares conceptual grounding with linkography [28], which visualizes how design ideas develop over time by tracing connections between moves. In this vein, techniques such as DraftMarks could evolve towards “forward-looking” linkography, visualizing emerging conceptual connections and gaps in human-ai collaborative writing. A potential direction for future work exploring generative linkography to support collaborative writing and enable new forms of epistemic feedback.

8.3 Skeuomorphic Signals and Interpretation

Skeuomorphic visual metaphors, as a design choice, come with both strengths and limitations. Skeuomorphism can leverage familiar physical affordances to make abstract processes like human-AI interaction visually intuitive and contextually embedded [43]. For writers and readers accustomed to pencil-and-paper revision practices, these metaphors evoke meaningful analogies, such as eraser marks suggesting revision and uncertainty. Compared to alternative techniques for process visibility (see Section 2), DraftMarks’ strength is that it integrates process cues into the text itself over detached metadata or summary metrics. However, these cues are not universally legible. For writers primarily exposed to digital environments or in contexts where physical revision artifacts are less common, skeuomorphic signals can lack resonance or even add confusion [72]. Future work can explore other metaphors within the space of typography, such as new fonts, the use of dynamic underlines, and color gradients, to represent process data in a more culturally neutral or scalable form. However, much like the floppy disk icon that is still widely used to represent “save” functionality, despite many users never having seen or used one. Skeuomorphic cues can persist even as their original referents lose relevance, raising questions about the intuitive nature of such metaphors across generations and cultures.

8.4 Stakeholder-Specific Design Considerations

Our formative evaluation revealed striking differences in stakeholder needs for process transparency. Teachers valued detailed collaboration insights for formative assessment, while academic reviewers viewed process visualization as unnecessary “extra work” that detracted from their efficiency-focused evaluation goals. General readers fell between these extremes, using transparency signals to assess authenticity and decide whether content warranted their attention. These findings suggest DraftMarks requires stakeholder-specific modes rather than a universal interface. An “instructor mode” might expose comprehensive process data, while a “publication mode” could focus on high-level authenticity signals. To realize this, future work should investigate configurable controller algorithms that flexibly govern which dimensions of transparency are emphasized, suppressed, or abstracted for different audiences.

However, this raises questions about who controls visibility settings when documents serve multiple audiences simultaneously. Designing flexible but principled controller algorithms thus requires not only technical advances in interface adaptability but also **normative frameworks** to arbitrate control, protect against selective disclosure, and uphold shared standards of writer agency and authenticity [1, 2, 5].

8.5 Limitations of Process Transparency: The Missing Citation Layer

While DraftMarks surfaces human–AI collaboration patterns, it does not address AI training data provenance. Our tool reveals when AI suggestions were integrated into the writing process, but it cannot expose the specific sources that informed those AI generations. This limitation is significant given growing concerns about AI systems reproducing copyrighted content without attribution, either verbatim or through close paraphrase. Such outputs risk crossing the line into plagiarism, especially when users unknowingly incorporate AI-generated text that echoes protected works [36]. Beyond individual academic integrity, this raises broader copyright challenges: without visibility into provenance, it is impossible to determine whether AI contributions constitute fair use, derivative works, or potential infringement. Future process-visualization tools should therefore not only trace collaboration patterns but also provide mechanisms for AI citation transparency, making visible when generated text may carry obligations of attribution. Realizing this vision, however, requires advances in AI explainability and provenance-tracking that extend well beyond current technical capabilities [26, 35].

9 Conclusion

As AI becomes a coauthor in everyday writing, understanding the creative processes by which a text was constructed has become essential for meaningful interpretation, evaluation, and learning. DraftMarks introduces a new form of process-aware signaling by embedding skeuomorphic visual cues directly into the text, helping readers see where and how AI was involved and where human effort was concentrated. By shifting process metadata from the margins to the foreground, DraftMarks provides transparency, encourages writer reflection, and can facilitate more grounded feedback from educators or peers. Our evaluation demonstrates DraftMarks’ affordances in helping readers better understand and assess human effort in writing, showing that in-text visualizations can meaningfully enhance how AI-assisted work is interpreted across educational and academic contexts.

References

- [1] Sanad Aburass and Maha Abu Rumman. 2024. Authenticity in Authorship: The Writer’s Integrity Framework for Verifying Human-Generated Text. *ArXiv* abs/2404.10781 (2024). <https://api.semanticscholar.org/CorpusID:269188035>
- [2] Roi Alfassi, Angel Cooper, Zoe Mitchell, Mary Calabro, Orit Shaer, and Osnat Mokry. 2025. Fanfiction in the Age of AI: Community Perspectives on Creativity, Authenticity and Adoption. *ArXiv* abs/2506.18706 (2025). <https://api.semanticscholar.org/CorpusID:279999533>
- [3] Ahmed S. Bahammam, Khaled Trabelsi, Seithikurippu R. Pandi-Perumal, and Hiatham Jahrami. 2023. Adapting to the Impact of AI in Scientific Writing: Balancing Benefits and Drawbacks while Developing Policies and Regulations. <https://api.semanticscholar.org/CorpusID:259138789>
- [4] Jeremy Birnholtz and Steven Ibara. 2012. Tracking changes in collaborative writing: edits, visibility and group maintenance. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) (CSCW ’12). Association for Computing Machinery, New York, NY, USA, 809–818. <https://doi.org/10.1145/2145204.2145325>
- [5] Jeremy P. Birnholtz and Steven Ibara. 2012. Tracking changes in collaborative writing: edits, visibility and group maintenance. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012). <https://api.semanticscholar.org/CorpusID:13524829>
- [6] Elizabeth L Bjork, Robert A Bjork, et al. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society* 2, 59–68 (2011), 56–64.
- [7] Alan F. Blackwell. 2006. The reification of metaphor as a design tool. *ACM Trans. Comput.-Hum. Interact.* 13, 4 (Dec. 2006), 490–530. <https://doi.org/10.1145/1188816.1188820>
- [8] ACM Publications Board. 2023. ACM Policy on Authorship. <https://www.acm.org/publications/policies/new-acm-policy-on-authorship>
- [9] Emma Bowman. 2022. A new AI chatbot might do your homework for you. But it’s still not an A+ student. *NPR* (2022). <https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia>
- [10] Richard Brath. 2021. *Visualizing with Text* (1st ed.). CRC Press, Taylor & Francis Group.
- [11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/147808706qp063oa>
- [12] Brisk Teaching. 2025. Inspect Writing: AI Tools for Inspecting, Grading, and Evaluating Student Work. <https://www.briskteaching.com/inspect-writing> Web-based educational tool for writing assessment and student progress tracking.
- [13] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems. *ArXiv* abs/2104.00358 (2021). <https://api.semanticscholar.org/CorpusID:232478535>
- [14] Liuqing Chen, Qianzhi Jing, Yixin Tsang, Qianyi Wang, Ruocong Liu, Duowei Xia, Yunzhan Zhou, and Lingyun Sun. 2024. AutoSpark: Supporting Automobile Appearance Design Ideation with Kansei Engineering and Generative AI. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) (UIST ’24). Association for Computing Machinery, New York, NY, USA, Article 108, 19 pages. <https://doi.org/10.1145/3654777.3676337>
- [15] Yuexi Chen, Yimin Xiao, Kazi Tasnim Zinat, Naomi Yamashita, Ge Gao, and Zhicheng Liu. 2025. Comparing Native and Non-native English Speakers’ Behaviors in Collaborative Writing through Visual Analytics. *arXiv preprint arXiv:2502.18681* (2025).
- [16] Zixin Chen, Jiachen Wang, Meng Xia, Kento Shigyo, Dingdong Liu, Rong Zhang, and Huamin Qu. 2025. StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 908–918. <https://doi.org/10.1109/TVCG.2024.3456363>
- [17] Fanny Chevalier, Pierre Dragicevic, Anastasia Bezerianos, and Jean-Daniel Fekete. 2010. Using text animated transitions to support navigation in document histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI ’10). Association for Computing Machinery, New York, NY, USA, 683–692. <https://doi.org/10.1145/1753326.1753427>
- [18] Won Ik Cho, Eunjung Cho, and Kyunghyun Cho. 2023. PaperCard for Reporting Machine Assistance in Academic Writing. *ArXiv* abs/2310.04824 (2023). <https://api.semanticscholar.org/CorpusID:263829109>
- [19] RD Coyne. 1992. The role of metaphor in understanding computers in design. *Proceedings of ACADIA (Association for Computer Aided Design in Architecture)* (1992), 3–11.
- [20] Susan D’Agostino. 2023. Academic experts offer advice on CHAT-GPT. <https://www.insidehighered.com/news/2023/01/12/academic-experts-offer-advice-chatgpt>
- [21] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2023. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Transactions on Computer-Human Interaction* 31 (2023), 1 – 40. <https://api.semanticscholar.org/CorpusID:266336118>
- [22] Douglas Engelbart and Jeff Rulifson. 1999. Bootstrapping our collective intelligence. *ACM Computing Surveys (CSUR)* 31, 4es (1999), 38–es.
- [23] Douglas C Engelbart. 2021. Augmenting human intellect: a conceptual framework (1962). (2021).
- [24] Center for Teaching Excellence. 2024. Ethical use of AI in writing assignments. <https://cte.ku.edu/ethical-use-ai-writing-assignments>
- [25] Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmquist, Tom Ewing, Keith N. Hampton, and Naren Ramakrishnan. 2015. ThemeDelta: Dynamic Segmentations over Temporal Topic Models. *IEEE Transactions on Visualization and Computer Graphics* 21, 5 (May 2015), 672–685. <https://doi.org/10.1109/TVCG.2014.2388208>
- [26] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627* (2023).

- [27] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) (*DIS '22*). Association for Computing Machinery, New York, NY, USA, 1002–1019. <https://doi.org/10.1145/3532106.3533533>
- [28] Gabriela Goldschmidt. 2014. *Linkography: unfolding the design process*. Mit Press.
- [29] Google for Developers. 2025. Structure of a Google Docs document. <https://developers.google.com/workspace/docs/api/concepts/structure> Google Workspace Docs API documentation.
- [30] Authors Guild. 2023. AG introduces New Publishing Agreement Clauses Concerning AI. <https://authorsguild.org/news/ag-introduces-new-publishing-agreement-clauses-concerning-ai/>
- [31] Alicia Guo, Shreya Sathyaranayanan, Lejje Wang, Jeffrey Heer, and Amy X. Zhang. 2024. From Pen to Prompt: How Creative Writers Integrate AI into their Writing Practice. *Proceedings of the 2025 Conference on Creativity and Cognition* (2024). <https://api.semanticscholar.org/CorpusID:273821820>
- [32] Marjin Haverbeke. 2025. ProseMirror. <https://prosemirror.net/> A toolkit for building rich-text editors on the web.
- [33] John R Hayes and Linda S Flower. 2016. Identifying the organization of writing processes. In *Cognitive processes in writing*. Routledge, 3–30.
- [34] Md Naimul Hoque, Tasnia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmquist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 1045, 15 pages. <https://doi.org/10.1145/3613904.3641895>
- [35] Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter* 26, 2 (2025), 21–43.
- [36] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185* (2023).
- [37] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. <https://doi.org/10.1145/3544548.3581196>
- [38] Adeeb M Jarrah, Yousef Wardat, and Patricia Fidalgo. 2023. Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies* (2023). <https://api.semanticscholar.org/CorpusID:260860417>
- [39] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST '23*). Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. <https://doi.org/10.1145/3586183.3606737>
- [40] Minsun Kim, SeonGyeom Kim, Suyoung Lee, Yooseang Yoon, Junho Myung, Haneul Yoo, Hyunseung Lim, Jeun Han, Yoonsoo Kim, So-Yeon Ahn, et al. 2024. Designing Prompt Analytics Dashboards to Analyze Student-ChatGPT Interactions in EFL Writing. *arXiv preprint arXiv:2405.19691* (2024).
- [41] Mina Lee, Katy Ilonka Gero, John Joot Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad Al Aghamdi, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–35.
- [42] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–19.
- [43] Sueyoon Lee, Abdallah El Ali, Maarten Wijntjes, and Pablo Cesar. 2022. Understanding and Designing Avatar Biosignal Visualizations for Social Virtual Reality Entertainment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 425, 15 pages. <https://doi.org/10.1145/3491102.3517451>
- [44] Haotian Li, Yun Wang, and Huamin Qu. 2024. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 845, 19 pages. <https://doi.org/10.1145/3613904.3642726>
- [45] Meilu Liu, Lawrence Jun Zhang, and Christine Biebricher. 2024. Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education* 211 (2024), 104977.
- [46] Robert F Lorch. 1989. Text-signaling devices and their effects on reading and memory processes. *Educational psychology review* 1 (1989), 209–234.
- [47] Brian J. McNely, Paul Gestwicki, J. Holden Hill, Philip Parli-Horne, and Erika Johnson. 2012. Learning analytics for collaborative writing: a prototype and case study. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (Vancouver, British Columbia, Canada) (*LAK '12*). Association for Computing Machinery, New York, NY, USA, 222–225. <https://doi.org/10.1145/2330601.2330654>
- [48] Meta. 2025. Lexical. <https://lexical.dev/> An extensible text editor framework that provides excellent reliability, accessibility and performance.
- [49] Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar A. Garcia Valencia, Fawad Qureshi, and Wisit Cheungpasitporn. 2023. Ethical Dilemmas in Using AI for Academic Writing and an Example Framework for Peer Review in Nephrology Academia: A Narrative Review. *Clinics and Practice* 14 (2023), 89–105. <https://api.semanticscholar.org/CorpusID:266699098>
- [50] Hannah Mieczkowski and Jeffrey Hancock. 2022. Examining Agency, Expertise, and Roles of AI Systems in AI-Mediated Communication. <https://doi.org/10.31219/osf.io/asny4>
- [51] Ayat Najjar, Huthaifa I. Ashqar, Omar A. Darwish, and Eman Hammad. 2025. Leveraging Explainable AI for LLM Text Attribution: Differentiating Human-Written and Multiple LLMs-Generated Text. *ArXiv abs/2501.03212* (2025). <https://api.semanticscholar.org/CorpusID:275337238>
- [52] Phong H. Nguyen, Kai Xu, Ashley Wheat, B.L. William Wong, Simon Attfield, and Bob Fields. 2016. SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 41–50. <https://doi.org/10.1109/TVCG.2015.2467611>
- [53] Donald A Norman. 2014. *The design of everyday things*. Mit Press.
- [54] United States Copyright Office. 2023. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence. https://copyright.gov/ai/ai_policy_guidance.pdf
- [55] OpenAI. 2022. ChatGPT. <https://chat.openai.com>
- [56] Purvish M. Parikh, Dinesh M. Shah, Urvish G. Parikh, Ajit Venkyoor, Govind Babu, Apurva Garg, and Hemant Malhotra. 2023. ChatGPT—Preliminary Overview with Implications for Medicine and Oncology. *Indian Journal of Medical and Paediatric Oncology* 44 (2023), 377 – 383. <https://api.semanticscholar.org/CorpusID:259155651>
- [57] Seong Ho Park. 2023. Authorship Policy of the Korean Journal of Radiology Regarding Artificial Intelligence Large Language Models Such as ChatGTP. *Korean Journal of Radiology* 24 (2023), 171 – 172. <https://api.semanticscholar.org/CorpusID:256758015>
- [58] Nitin Liladhar Rane. 2024. Enhancing the quality of teaching and learning through ChatGPT and similar large language models: Challenges, future prospects, and ethical considerations in education. *TESOL and Technology Studies* (2024). <https://api.semanticscholar.org/CorpusID:267951872>
- [59] Revision History. 2025. Revision History: Track Student Writing in Google Docs. <https://www.revisionhistory.com/> Chrome extension for analyzing student writing patterns and academic integrity in Google Docs.
- [60] 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.
- [61] Jen Rogers and Anamaria Crisan. 2023. Tracing and Visualizing Human-ML/AI Collaborative Processes through Artifacts of Data Work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 837, 22 pages. <https://doi.org/10.1145/3544548.3580819>
- [62] Marlene Scardamalia and Carl Bereiter. 1987. Knowledge telling and knowledge transforming in written composition. *Advances in applied psycholinguistics* 2 (1987), 142–175.
- [63] Antoinette Shibani, Simon Knight, Kirsty Kitto, Ajani Karunanayake, and Simon Buckingham Shum. 2024. Untangling Critical Interaction with AI in Students' Written Assessment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI EA '24*). Association for Computing Machinery, New York, NY, USA, Article 357, 6 pages. <https://doi.org/10.1145/3613905.3651083>
- [64] Antoinette Shibani, Ratnavel Rajalakshmi, Faeria Mattins, Srivarshan Selvaraj, and Simon Knight. 2023. Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing. In *Proceedings of the 16th International Conference on Educational Data Mining*, Mingyu Feng, Tanja Käser, and Partha Talukdar (Eds.), International Educational Data Mining Society, Bengaluru, India, 183–193. <https://doi.org/10.5281/zenodo.8115695>
- [65] Momin Siddiqui, Roy Pea, and Hari Subramonyam. 2025. Script&Shift: A Layered Interface Paradigm for Integrating Content Development and Rhetorical Strategy with LLM Writing Assistants. *arXiv preprint arXiv:2502.10638* (2025).
- [66] Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. 2023. Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions. *ArXiv abs/2307.14107* (2023). <https://api.semanticscholar.org/CorpusID:260164854>

- [67] F. Sperre, H. Schäfer, D. Keim, and M. El-Assady. 2021. Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement. *Computer Graphics Forum* 40, 3 (2021), 215–226. <https://doi.org/10.1111/cgf.14301> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14301>
- [68] Sangho Suh, Bryan Min, Srishti Palani, and Hajun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST '23*). Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. <https://doi.org/10.1145/3586183.3606756>
- [69] Chengzheng Sun and Clarence Ellis. 1998. Operational transformation in real-time group editors: issues, algorithms, and achievements. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) (*CSCW '98*). Association for Computing Machinery, New York, NY, USA, 59–68. <https://doi.org/10.1145/289444.289469>
- [70] Tiptap. 2025. Tiptap - Dev Toolkit Editor Suite. <https://tiptap.dev/> The headless and open source editor framework.
- [71] Johnny Torres, Sixto García, and Enrique Peláez. 2019. Visualizing authorship and contribution of collaborative writing in e-learning environments. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) (*IUI '19*). Association for Computing Machinery, New York, NY, USA, 324–328. <https://doi.org/10.1145/3301275.3302328>
- [72] Inês Cunha Vaz Pereira Urbano, João Pedro Vieira Guerreiro, and Hugo Miguel Aleix Albuquerque Nicolau and. 2022. From skeuomorphism to flat design: age-related differences in performance and aesthetic perceptions. *Behaviour & Information Technology* 41, 3 (2022), 452–467. <https://doi.org/10.1080/0144929X.2020.1814867> arXiv:<https://doi.org/10.1080/0144929X.2020.1814867>
- [73] Frank van Ham, Martin Wattenberg, and Fernanda B. Viégas. 2009. Mapping Text with Phrase Nets. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1169–1176. <https://doi.org/10.1109/TVCG.2009.165>
- [74] Fernanda B. Viégas and Martin Wattenberg. 2008. TIMELINES Tag clouds and the case for vernacular visualization. *Interactions* 15, 4 (July 2008), 49–52. <https://doi.org/10.1145/1374489.1374501>
- [75] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) (*CHI '04*). Association for Computing Machinery, New York, NY, USA, 575–582. <https://doi.org/10.1145/985692.985765>
- [76] Fernanda B. Viégas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137–1144. <https://doi.org/10.1109/TVCG.2009.171>
- [77] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI EA '20*). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3334480.3381069>
- [78] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI '15*). Association for Computing Machinery, New York, NY, USA, 1865–1874. <https://doi.org/10.1145/2702123.2702517>
- [79] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1221–1228. <https://doi.org/10.1109/TVCG.2008.172>
- [80] Laura Weidinger, John Mellor, Maribeth Rauth, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glæse, Børja Balle, Atoosha Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
- [81] J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Visualization 1995 Conference*. 51–58. <https://doi.org/10.1109/INFVIS.1995.528686>
- [82] Zhihao Yao, Yao Lu, Yijie Guo, Qirui Sun, Mingyue Gao, and Haipeng Mi. 2024. Taoie: Designing a Museum Robot utilizing Cultural Metaphors. In *Proceedings of the Eleventh International Symposium of Chinese CHI* (Denpasar, Bali, Indonesia) (*CHCHI '23*). Association for Computing Machinery, New York, NY, USA, 241–250. <https://doi.org/10.1145/3629606.3629628>
- [83] Haneul Yoo, Junho Myung, Hwajung Hong, Yoo Lae Kim, So-Yeon Ahn, Minsun Kim, Jieun Han, Juho Kim, Tak Yeon Lee, Hyunseung Lim, and Alice H. Oh. 2023. RECIPE: How to Integrate ChatGPT into EFL Writing Education. *Proceedings of the Tenth ACM Conference on Learning @ Scale* (2023). <https://api.semanticscholar.org/CorpusId:258823196>
- [84] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST '23*). Association for Computing Machinery, New York, NY, USA, Article 5, 30 pages. <https://doi.org/10.1145/3586183.3606800>