SpatialBalancing: Designing an LLM-Powered Spatial Externalization Interface for Iterative Science Communication Writing

ANONYMOUS AUTHOR(S)

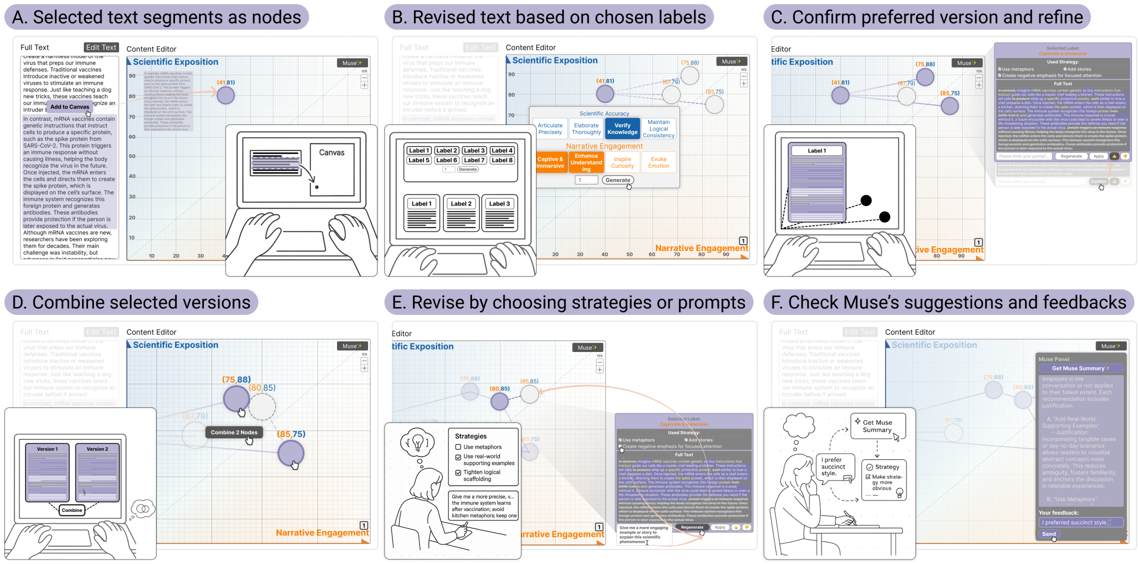

Fig. 1. Example Workflow of using SpatialBalancing for iterative science communication writing. A – Jenny drags her draft into the canvas, where each paragraph becomes a node mapped by Scientific Exposition (Y-axis) and Narrative Engagement (X-axis). B – She selects revision labels such as Enhance Understanding or Captivate & Immerse, each tied to LLM-driven strategies that generate new versions placed accordingly. C – Jenny reviews and confirms preferred revisions, which turn purple for further refinement. D – She can combine two versions into a synthesized draft, balancing credibility and engagement. E – Further revisions are guided by strategies or custom prompts, enabling precise, iterative control. F – Finally, SpatialBalancing's Muse assistant reflects on her revision history and offers adaptive suggestions.

Revising science communication is inherently challenging: writers must iteratively balance scientific exposition and narrative engagement, often drifting back and forth between these competing directions. While prior HCI systems have made LLM-assisted writing more accessible, they offer limited help for navigating this kind of cumulative, multi-directional revision process. In this work, we frame science communication revision as movement within a two-dimensional rhetorical space and present SpatialBalancing, an exploratory interface that externalizes goals, revision states, and trajectories through spatial visualization. By constructing a design space of communication strategies and embedding them into a spatial exploratory canvas, our system treats feedback as navigational cues rather than prescriptive judgments. Our findings show that spatial externalization helps writers stay oriented to goals, reason about revision as a trajectory, and explore alternatives at low cost, supporting greater metacognitive control and confidence without increasing workload. Together, this work highlights how spatial externalization can reframe LLM-assisted revision from producing better text to supporting better thinking over time.

Manuscript submitted to ACM

CCS Concepts: • **Human-centered computing → Collaborative interaction**.

Additional Key Words and Phrases: Narrative Strategy, Science Communication, Writing Assistance, Human-AI collaboration

ACM Reference Format:

Anonymous Author(s). 2026. SpatialBalancing: Designing an LLM-Powered Spatial Externalization Interface for Iterative Science Communication Writing. In *CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/3474349.3480203

1 Introduction

Writing is fundamentally a non-linear process of knowledge transformation, requiring writers to cycle recursively through planning, translating, and reviewing rather than producing a linear output [23, 63]. Throughout this process, writers must balance multiple rhetorical goals, making local revisions while maintaining global coherence [43, 57]. Recent advances in large language models (LLMs) have lowered the cost of generating and revising text at scale [43, 63]. In response, many HCI systems operationalize specific rhetorical strategies or scaffold discrete aspects of drafting and rewriting [36, 65, 75, 76]. Others support non-linear exploration by helping writers generate, compare, and organize multiple text variations [57, 63, 79], or by breaking down or re-organize feedback to make revision more actionable [57, 63, 75, 76]. However, these approaches primarily externalize the products of revision, while leaving the rhetorical goal space that guides revision decisions implicit. As a result, writers must internally reason about how successive revisions advance or compromise competing goals [42, 77], making revision cognitively demanding, particularly in complex knowledge domains such as science communication [72, 77].

Science communication writing differs fundamentally from academic prose. Rather than focusing solely on exposition whose purpose is to convey relevant facts and knowledge, it must translate complex knowledge into forms that are understandable and memorable for non-expert audiences [7, 33, 53]. Narrative techniques such as storytelling, metaphor, and suspense are widely used to achieve this goal, as they can increase attention and comprehension and make the content more engaging [26]. However, narrative also introduces persistent tension: emphasizing entertainment risks oversimplification or loss of credibility [18, 26, 50]. On the other hand, overly technical or serious exposition can alienate non-expert readers by demanding sustained cognitive effort while offering few cues for relevance or engagement [7, 13, 17]. Effective science communication, therefore, requires continual balancing between scientific exposition and narrative engagement, which is an inherently iterative process, where writers repeatedly revise and reassess the two rhetorical goals to reach a sweet point rather than making a single stylistic decision upfront [13, 26, 30, 77]. As science content proliferates across platforms like YouTube and TikTok, a growing number of "everyday" creators, many lacking formal communication training, are taking on the role of science explainers [47]. These creators increasingly turn to LLM tools to support ideation, drafting, and real-time feedback throughout the revision process [48]. This shift furthure amplifies the need for interfaces that provide comprehensive guidance to support the iterative work of balancing exposition and engagement across multiple revision cycles.

To address this gap, **we explore how interface design can better support goal-aware, iterative revision in science communication writing for non-expert creators.** Building on prior work in science communication writing, we construct a design space of rhetorical strategies for enhancing scientific exposition and narrative engagement, and use it to inform an initial prototype. Through this process, we identify the limitations of strategy-centric and linear revision workflows, motivating a shift toward externalizing revision goals and trajectories.

One promising direction is **spatial externalization**. Recent canvas-based interfaces like PatchView [11] and Luminate [66] demonstrate how spatial layouts enable users to navigate and compare LLM outputs, yet they primarily externalize LLM generated content attributes rather than the revision goals and trajectories that guide iterative writing decisions. Building on this line of research and theories of *Thinking with external representations* [37], which posit that making cognitive structures visible in the environment reduces the cost of tracking change and enables more deliberate exploration, we present SpatialBalancing, an LLM-powered spatial interface that reframes revision as navigation through a two-dimensional rhetorical space. The system externalizes scientific exposition and narrative engagement as persistent spatial dimensions, enabling writers to visualize where each revision stands and track how successive edits shape the draft over time. Rather than offering prescriptive judgments, SpatialBalancing provides feedback as navigational cues that support low-cost exploration, comparison, and reflection across multiple revision directions.

A controlled user study demonstrates that spatial externalization helps writers maintain orientation toward rhetorical goals, conceptualize revision as a trajectory, and exercise greater metacognitive control. Simultaneously, our findings surface important tensions such as over-reliance on externalized guidance, which may encourage metacognitive laziness, these insights pointing toward critical design opportunities for future LLM-assisted revision interfaces. This work contributes to the field in the following ways:

(1) A constructed design space of 25 science communication strategies organized into eight actionable labels that operationalize scientific exposition and narrative engagement for LLM-based revision support.

(2) SpatialBalancing: A spatial externalization interface for goal-aware LLM-assisted revision that externalizes rhetorical goals and revision trajectories through a 2D exploratory canvas, enabling writers to navigate different objectives and maintain metacognitive control across iterations.

(3) Design insights for future LLM-assisted writing interfaces derived from iterative design and user study evaluation, pointing to the importance of mitigating over-reliance on externalized feedback, preserving user agency through adaptive externalization, and providing embedded reflective support throughout the revision process.

2 Related Work

2.1 Balancing Scientific Exposition and Narrative Engagement in Science Communication Writing

In the Information Age, online science communication has become increasingly dominant, especially in the popular science field [9, 51]. Science communication refers to the strategic use of various forms of communication, such as media, events, and interactions, to convey scientific information to diverse audiences in a way that aims to increase awareness, enjoyment, interest, opinion-forming, and understanding [7, 33, 53]. The popular science movement (also known as pop science or popsci) aims to interpret and present scientific concepts in an accessible way for a general audience, placing greater emphasis on entertainment and broadening its scope compared to traditional science journalism [5, 15, 71]. As online communication technologies have become more accessible, various formats have emerged to deliver popular science content, including books, documentaries, web articles, and online videos [21, 71, 78].

A fundamental challenge in science communication writing lies in balancing two often competing dimensions: scientific exposition and narrative engagement [18, 26, 50]. Expository writing applies to tasks whose purpose is to convey relevant facts and knowledge, while narrative applies to tasks whose goal is to convey

an account of real through telling a story [43]. Burns et al. [7] made a vivid analogy, describing science communication writing as a form of "mountain climbing," balancing between scientific literacy and science culture. Similarly, Dahlstrom [13] emphasized that science communication writing inherently involves both narrative and expository elements. In this study, we use the terms "scientific exposition" and "narrative engagement" to describe this tradeoff [17], because these terms more directly capture the practical tension between maintaining rigorous, detailed scientific facts presentation and creating compelling, accessible content for diverse audiences [17, 50]. In practice, achieving this balance is inherently iterative rather than a one-shot optimization. If a draft over-indexes on expository, logical-scientific presentation, it may preserve accuracy but often becomes harder for non-experts to process and remember; narrative formats [13, 26]. At the same time, leaning too far toward narrative can create a different failure mode: narratives are intrinsically persuasive and are often evaluated by verisimilitude (how “true-to-life” they feel) rather than the accuracy standards of logical-scientific discourse, which raises ethical and credibility risks in science communication [13, 30]. Consequently, writers must revise through multiple passes—adjusting where and how narrative devices are woven into explanatory content, because the effectiveness of a change depends on its relationship to the surrounding narrative structure and the reader’s evolving interpretation, not just the local wording [26].

The tension between these dimensions stems from their fundamentally different linguistic requirements. Engaging content relies on narrative techniques—storytelling, analogy, and suspense to capture attention [13, 21, 26], while scientifically content demands rigorous expository writing that prioritizes scientific detail and credibility [35, 38]. Recent HCI systems have begun operationalizing specific strategies within LLM-powered co-creation tools to lower the barrier for science communication writing, particularly for non-expert writers who constitute the dominant group on online platforms. For example, systems such as Metaphorian support metaphor creation through LLM-assisted exploration [36, 77], while AI workflows for Tweetorials scaffold the generation of hooks, examples, and anecdotes to engage general audiences [46, 77]. However, these systems typically focus on supporting the application of individual strategies at specific moments in writing, rather than the broader iterative revision process in which writers must continuously rebalance scientific exposition and narrative engagement. To mimic this gap, in this study, we design an LLM-powered visualization interface to support the iterative revision process of balancing scientific exposition and narrative engagement, grounded in a holistic understanding of communication strategies for achieving this balance.

2.2 Iterative Revision through Co-creation with LLM

Prior work has characterized writing as an inherently iterative process involving distinct stages, such as revision, and has emphasized that writing tasks are driven by multiple rhetorical purposes rather than a single objective [43]. These purposes, including expository, narrative, persuasive, and educational goals, often coexist and shape revision decisions in audience-dependent ways [43]. Iterative revision toward multiple rhetorical goals remains hard because writers must repeatedly shift attention across levels and keep track of what changed, why it changed, and which direction each revision moves the draft [63, 75, 76]. Prior work shows that today’s dominant linear document interfaces with chat functions still constrain this kind of non-linear goal juggling: prompting across micro/macro levels requires manual cross-referencing and repeated prompt formulation, which disrupts writers’ flow and makes it difficult to sustain coherent rhetorical strategy across iterations [65]. HCI systems have begun addressing parts of this problem by externalizing revision materials in more navigable forms. ABScribe, for example, tackles the “too-many-variants” problem by helping writers create, compare, and revise multiple text variations without overwriting or clutter, explicitly aligning LLM use with revision’s recursive, non-linear nature [57]. Friction scaffolds reflection by breaking feedback into actionable units and guiding iterative revision cycles [75], while Synthia uses visual organization plus traceable links among feedback, source text, and revisions to support non-linear branching and exploration rather than one-shot rewriting [76]. However, existing systems primarily externalize revision artifacts, such as alternative drafts, layers, or feedback, rather than the rhetorical goal space that guides revision decisions. As a result, writers must internally reason about how successive edits advance or compromise competing goals, making trade-offs opaque and cognitively demanding and leading to trial-and-error prompting [65], especially in complex domains like science communication.

One promising direction for addressing these challenges is externalization—using visualization to make the exploratory space of revision perceptible and navigable, which has long been shown to support complex reasoning by offloading, structuring. According to the theory of *Thinking with external representations*, making goals, states, and relations perceptible in the environment can reduce the cognitive cost of tracking change, support orientation, and enable more deliberate exploration of alternatives [37]. Building on this insight, prior HCI systems have leveraged visualization and spatial exploration to externalize latent aspects of generative processes with LLM. PatchView [11] and Luminate [66] organize LLM outputs within navigable visual spaces to support sensemaking, comparison, and steering, while Toyteller [12] shows how visual manipulation can function as an expressive control channel for generative storytelling. These systems demonstrate how spatial externalization can shift generative interaction from linear prompting toward structured exploration over a visualized space of possibilities. However, this line of work externalizes content attributes or generative alternatives, while leaving the iterative revision process characterized by sustained goal juggling, cumulative decision-making, and cross-iteration reasoning largely unsupported. Our work builds on this line of research by applying externalization to construct an exploratory space that makes rhetorical goals and revision trajectories explicit through visualization, supporting goal-oriented iterative revision in science communication writing.

3 Iterative User-Centred Design

3.1 Design Space Construction of Science Communication Strategies

Science communication writing involves balancing multiple rhetorical goals, most notably accurate scientific exposition and engaging narrative expression [18, 26, 50]. Writers achieve different balances by applying a diverse set of rhetorical strategies, often adapting their choices based on audience characteristics and communicative intent [77]. Although prior work in communication studies has identified a rich set of rhetorical strategies for science communication, aiming at capturing public attention, improving memorability [30, 78]. These strategies are rarely examined through a systematic lens that foregrounds the dual rhetorical goals of scientific rigorous expression and narrative engagement.

To support structured exploration and interaction design around rhetorical revision, it is therefore necessary to explicitly identify, organize, and formalize these strategies. Motivated by this need, we aimed to construct a design space of rhetorical strategies that support narrative engagement and scientific exposition.

To form the design space, we conducted a literature review in related fields, specifically in communication studies, education, psychology, linguistics and writing, and HCI, to identify writing strategies that can enhance narrative engagement and scientific exposition. We searched keywords "science communication" OR "scientific writing" OR "popular science" AND "strategy" OR "strategies" OR "method" in Google Scholar, the ACM Digital Library, and the IEEE Xplore Digital Library. Thus, we broaden our search to the discussion of the narrative or narrative design of learning content in general. We finally chose 35 papers across education (9), psychology (5), communication studies (15), and HCI (6) that are highly relevant to our research. They are chosen because they focus on methods and strategies for designing narratives that potentially improve knowledge rentention and create engaging narratives [21, 52]. Additionally, some of the papers explore related fields, such as the analysis of narrative peaks in data videos [73] or documentaries [41].

Two authors participated in the coding of these 35 papers. The primary objective was to identify potential peak narrative strategies for balancing scientific exposition and narrative engagement in these previous studies. Initially, each author independently reviewed all the selected papers, focusing on content related to narrative strategies or structures

Table 1. Design Space of Science Communication Writing Strategies.

Scientific Exposition			
Label 1 **Articulate Precisely** Communicates scientific concepts with exposition and clarity, using appropriate terminology and well-defined language to prevent ambiguity or misinterpretation [31, 35, 52].	**Label 2** **Elaborate Thoroughly** Provides sufficient detail or comprehensive theoretical discussion by unpacking underlying mechanisms, explaining implications, and citing evidence to elaborate on the knowledge point while avoiding bias [32, 39].	**Label 3** **Verify Knowledge** Supports claims with credible sources, data, or reasoning, allowing audiences to feel more trustworthy of the given information [39, 58].	**Label 4** **Maintain Logical Consistency** Ensures that arguments and explanations are coherent and internally consistent, following a clear logical structure [68].
Strategies: (4) Acknowledge Uncertainties, (5) Consistent Terminology, (18) Simplify and abstract language, (19) Clarify Key Terms, (21) Repeat key point(s) or question(s), (22) Emphasize with Numbers	**Strategies:** (3) Step-by-Step Explanation, (4) Acknowledge Uncertainties, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (25) Tie Science to Current Events	**Strategies:** (2) Rigorous Source Verification, (6) Citations & Quotes, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (7) Everyday Events to Scientific Insights Events	**Strategies:** (1) Layered Transitions, (3) Step-by-Step Explanation, (20) Key Point Recap, (23) Strengthen the Connections Between Content
Narrative Engagement			
Label 5 **Captivate & Immerse** Engages the audience's attention and draws them into the narrative or content flow by adding stories [26, 45] or using intriguing language [21, 52].	**Label 6** **Enhance Understanding** Help audiences to grasp complex scientific ideas using rational, structural content or vivid analogies, visualizations [21, 26, 30].	**Label 7** **Inspire Curiosity** Stimulates the audience's desire to learn more and have motivation to further explore by applying different forms of questions [40].	**Label 8** **Evoke Emotion** Creates an emotional response, positive or negative, and makes the audience feel connected to the content, even immerse themselves in the described scenario [26, 59].
Strategies: (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal, (11) Use metaphors, (12) Inject humor, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire	**Strategies:** (11) Use metaphors, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (21) Repeat key point(s) or question(s), (23) Strengthen the Connections Between Content, (24) Present Balanced Views, (25) Tie Science to Current Events	**Strategies:** (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal	**Strategies:** (9) Reflection Question, (12) Inject humor, (14) Add stories, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire, (21) Repeat key point(s) or question(s)

Note. Specific information about each strategy (e.g., definitions, examples) is presented in Table 5.

that enhance knowledge retention, recall, focus, or contribute to engagement and curiosity. Relevant content was then extracted and compiled into a consolidated document. Subsequently, using an open coding approach [3], two authors independently identified and coded key strategies, including their definitions and relevant contexts within the selected content. Following this, the two authors engaged in multiple discussion sessions to reconcile differences and reach a consensus on the coding. Finally, we identified a initial draft of strategies from these selected papers.

Then, we conducted a Focus Group Discussion (FGD) [55] with the four experts. Together, we refined our initial strategy design space by clarifying the definition and use of each strategy, and classified the communication strategies by their functions. In this design space, we categorized the 25 identified strategies into three groups: those that enhance narrative engagement (N=10), those that enhance scientific exposition (N=7), and those that enhance both (N=8). This process yielded four labels each for scientific exposition and narrative engagement. Some strategies, due to their multifunctionality, were assigned to multiple labels, forming the final design space (Table 1).

This design space provides a structured foundation for subsequent system design by externalizing rhetorical revision strategies as discrete, reusable units aligned with the two core rhetorical goals of science communication writing.

3.2 Initial Prototype and Iteration

Building on prior work that demonstrates how large language models can lower the barriers of science communication writing by operationalizing rhetorical strategies as generative and revisable resources [36, 46, 77], we draw on insights from the narrative design space of science communication to inform our design. We develop an initial prototype as a set of design probes to ground subsequent system design for supporting complex, multi-goal revision in LLM-assisted science communication writing.

3.2.1 Initial Prototype. (Figure 2) Our initial prototype was designed as a lightweight design probe, consisting of a basic text editor and a strategy selection panel. The panel presented all 25 identified strategies, organized under their corresponding labels derived from the design space. Selecting a label expanded the associated strategies, allowing users to browse and choose a specific rhetorical strategy.

In the strategy-specific mode, upon selecting a strategy, the system analyzed the textual context and highlighted candidate segments where the chosen strategy could be applied. By clicking on a highlighted segment, users could preview an LLM-generated revision that instantiated the selected strategy. In addition, in the content-specific mode, users could directly highlight a passage in the text, and the system would surface a set of recommended strategies relevant to that passage. Users could then preview alternative revisions generated using different strategies. When finishing an edit using a specific strategy in a specific passage, the system displayed supplementary explanations in a side panel, including a brief description of the selected strategy and the rationale for its application to the paragraph. After confirming a revision, users could further edit the text manually, retaining full control over the final outcome.

Behind the scenes, revisions were generated through a prompt-based LLM workflow grounded in the defined strategy descriptions, curated examples, and the surrounding textual context. Dedicated backend functions supported strategy recommendation in content-specific mode, target text selection in strategy-specific mode, and revision generation, enabling flexible and iterative interaction between users and the LLM.

3.2.2 Participants and Procedure. To elicit design insights from the initial prototype, we conducted a formative study with six participants recruited from the university community who had prior experience creating science communication content but not science communication writing experts. All participants were experienced writers and reported extensive prior use of LLM-based writing tools.

Each session began with a brief walkthrough of the prototype, during which we introduced the available interactions and strategy-based revision workflow. Participants were then asked to revise a short science communication text about déjà vu, adapted from publicly available reference material, into a version that was both using rigorous in scientific exposition and engaging for a general audience. We employed a think-aloud protocol, encouraging participants to verbalize their reasoning, challenges, and desired alternative functionalities as they interacted with the system. At the end of each session, participants reflected on their overall experience and provided suggestions for improvement in a semi-structured discussion. Each session lasted approximately 45 minutes.

3.2.3 Feedback and Design Consideration.

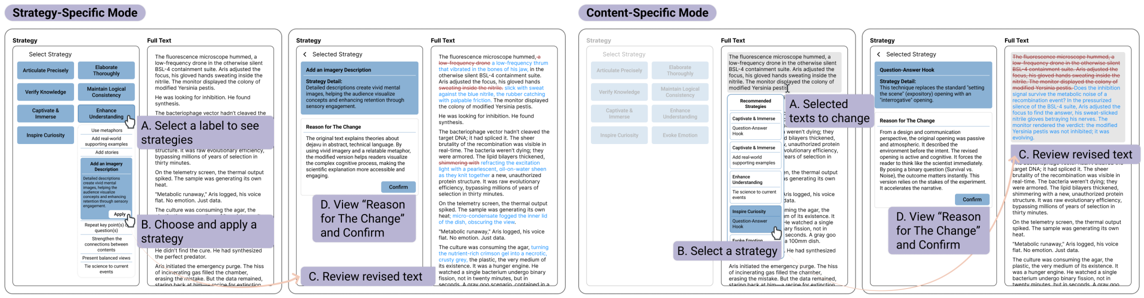

Fig. 2. The prototype consisted of a text editor and a strategy selection panel organized by the design space. Users could revise text through two interaction modes: strategy-specific, where selecting a rhetorical strategy highlighted candidate segments and previewed LLM-generated revisions, and content-specific, where selecting a text segment surfaced relevant strategies and alternative revisions. For each revision, the system provided a brief explanation of the applied strategy and its rationale. Users retained full control by confirming, rejecting, or manually editing revisions. All revisions were generated through a prompt-based LLM workflow grounded in strategy definitions, curated examples, and local textual context.

Lack of Continuous Goal Orientation . Participants viewed strategies as means toward higher-level communicative intentions, shaped by audience, platform, and purpose (P1,P2). Four out of six participants (P1, P4, P5, P6) expressed a desire for real-time feedback that reflects how their revisions might be interpreted by the target audience. As P1 explained, "although authors may intentionally adjust rhetorical strategies for different audiences—for instance, using more narrative elements for children — but they often lack visibility into how those audiences would actually respond, such as whether the revised content feels sufficiently engaging or easy to understand. " This highlights a need for system designs that provide real-time feedback during revision, enabling users to understand how their revisions are progressing towards editing goals.

Difficulty Reasoning About Cumulative Change. Participants consistently emphasized the need to track and reflect on their own revision trajectories, rather than treating revisions as isolated edits. P1, P3 and P6 wanted to see where changes occurred, which strategies were applied, and how these decisions accumulated over time. P2 expressed a desire to "track where I changed things so he can improve my own revision process." P3 further articulated a need for a timeline-based history, in which strategy selections, modified text spans, deleted context, and resulting versions could all be traced and revisited. These suggest that iterative science communication writing is not only about producing better text, but also about developing an understanding of one's own revision behavior over time. Participants expressed a desire for the system to capture revision history as a reflective artifact, making patterns of strategy use visible and supporting deliberate backtracking, comparison, and learning across revisions.

Cognitive Overload from Unstructured Strategy Presentation. While participants appreciated the richness of the strategy set, all of them found that presenting all strategies at once created cognitive overload. This overload manifested in three ways. First, learning burden. As P5 pointed out, "familiarizing oneself with all available strategies can be cognitively demanding. Writers tended to rely on familiar strategies, while unfamiliar ones incurred additional learning effort. Second, lack of structure. P2 suggested that strategies could be "packaged" according to different purposes, while P3 noted that the current interaction design made it difficult to compare options simultaneously. Third, absence of hierarchical guidance. P4 expressed a desire for more hierarchical guidance, such as high-level structural suggestions before more localized paragraph- or sentence-level recommendations. Together, these observations indicate that for non-expert writers, effective support lies not in maximizing choice, but in offering structured, context-sensitive strategy recommendations that lower cognitive load.

3.3 Design Goals

Based on the research gap from the literature review, insights from design space construction, and initial prototype iteration, we propose the following design goals:

Design Goal 1: Externalize Rhetorical Goals to Support Goal-Aware Iterative Revision Prior LLM-assisted writing systems embed rhetorical intentions implicitly through prompts or localized strategy use, requiring writers to internally track communicative goals across revisions [63, 65]. The system should externalize rhetorical goals as explicit, inspectable reference points, enabling writers to reason about revision directions and assess progress toward intended balances between scientific exposition and narrative engagement.

Design Goal 2: Represent Revision as a Trajectory Rather Than Isolated Edits or Alternatives Existing systems primarily support comparison among alternative drafts or localized revisions [57, 75, 76], offering limited support for understanding how revisions accumulate over time. The system should represent revision as a continuous, traceable trajectory that links versions, applied strategies, and resulting changes, supporting reflection, backtracking, and learning across iterative revisions.

Design Goal 3: Design an Exploratory Space to Gradually Guide Strategy Use While prior work operationalizes individual rhetorical strategies to lower barriers to science communication writing through co-creation with LLM [36, 46], exposing a large strategy space often overwhelms non-expert users and reinforces habitual choices. The system should design an exploratory space that gradually guides strategy use through structured, context-sensitive cues, supporting discovery and comparison over time while reducing cognitive burden and preserving user agency.

4 SpatialBalancing: An Spatial Externalized Visualization Interface for Navigable LLM-Assisted Revision

Grounded in our design goals, we design SpatialBalancing around *externalization*—making rhetorical goals, revision states, and their evolution visible and manipulable during writing. This choice is motivated by *Thinking with External Representations*, which argues that external representations can support complex reasoning by providing stable reference points for orientation, reducing internal tracking demands, and enabling deliberate exploration of alternatives [37].

We draw inspiration from canvas-based Spatial LLM interfaces such as PatchView [11] and Luminate [66], which show how spatial layouts and overview-to-detail navigation can support exploratory interaction with generated alternatives. Building on these interaction principles, SpatialBalancing uses an exploratory canvas not to organize content attributes or design variants, but to externalize rhetorical goals and revision trajectories, allowing writers to interpret progress as movement through a navigable space.

4.1 SpatialBalancing as an Exploratory Space

SpatialBalancing comprises a left-hand text editor and a right-hand exploratory canvas (Figure 4). users can send any span—sentence, paragraph, or full draft—to the canvas for iterative revision. Each version is plotted in a 2D space (x: Narrative Engagement; y: Scientific Exposition); gray points denote exploratory drafts and purple points mark confirmed selections, which can be further refined via labels or custom edits. This spatial view makes revision states and decision points explicit, helping users balance exposition and engagement.

The canvas supports branch-based exploration with three zoom levels (Figure 3). Dropped text becomes a root node; applying labels or custom instructions spawns child nodes, forming a tree that traces exploration paths. At 0–30% zoom,

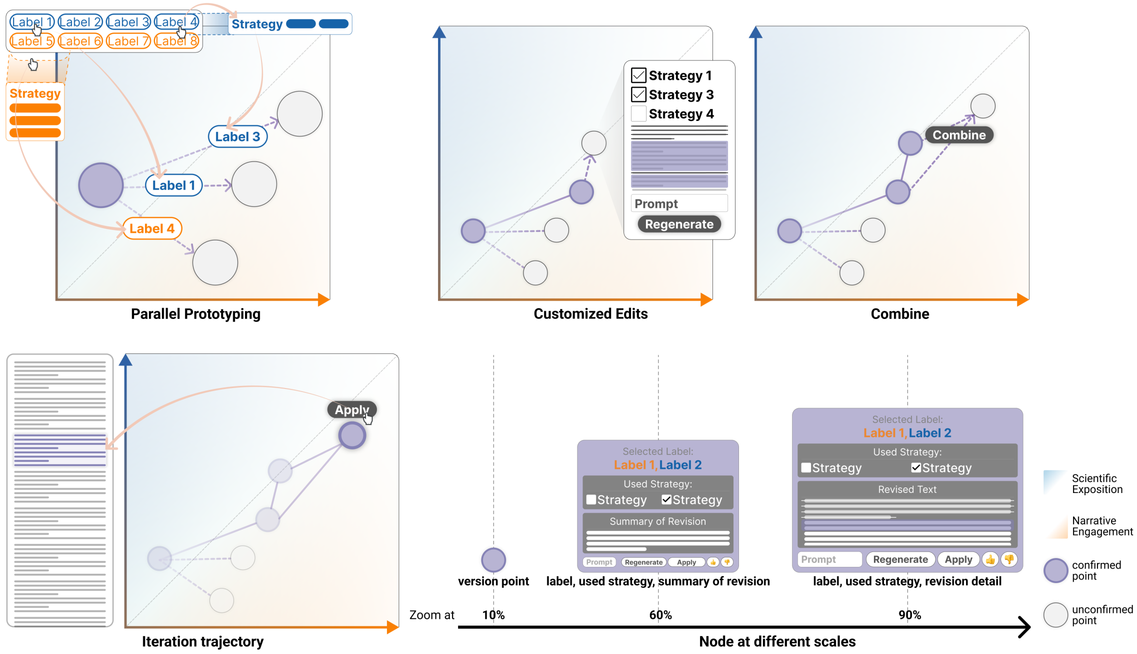

Fig. 3. (1) SpatialBalancing support parallel prototyping with diverse directions of LLM output; users can use customized edits like change specifc strategy and combine different LLM output to generate new nodes. The 2D coordinate space also allow user to see their iteration trajectory. (2) SpatialBalancing canvas supports three zoom levels: dots for version overview (0–30%), change summaries with labels and strategies (40–70%), and full content with highlights of edits (80–100%).

points provide an overview; at 40–70%, summaries show per-version changes and chosen strategies; at 80–100%, full text with diffs against the original is displayed. This progressive disclosure enables rapid comparison and reflective choice among alternatives.

4.2 Spatial Externalization Features to Support Goal-aware Revision

4.2.1 Real-Time Two-Axis Goal Externalization (DG1). To support goal-aware revision (DG1), SpatialBalancing externalizes rhetorical goals through real-time two-axis feedback. Each version of the text is represented as a point in a two-dimensional space, where one axis encodes narrative engagement and the other scientific exposition. This representation transforms abstract revision goals into stable, perceptible reference points, enabling users to orient themselves and reason about the direction of their revisions. Whenever users create or modify a version, a Scorer Agent (Explained in Section 4.4) assigns engagement and exposition scores based on audience-informed criteria, which determine the node's position on the canvas.

4.2.2 Strategy Recommendation via Rhetorical Labels (DG1 & DG3). (Figure 5(1)) To support goal-aware revision while managing cognitive load (DG1, DG3), SpatialBalancing introduces an eight-label taxonomy that scaffolds strategy exploration around two overarching rhetorical goals: scientific exposition and narrative engagement. Four labels guide revisions toward strengthening scientific explanation, while the other four foreground narrative techniques for engagement. Rather than requiring users to reason over individual strategies, these labels decompose abstract rhetorical goals into actionable revision directions. By selecting one or more labels aligned with their intentions, writers receive guided yet flexible revisions generated by the LLM, reducing the burden of exhaustive choice while providing clear direction for exploration.

4.3 Spatial Externalization Features to Enable Trajectory-Based Revision Reasoning

4.3.1 Fine-Grained Control for Specific Versions (DG3). (Figure 5(2)) To complement structured guidance with user control (DG3), SpatialBalancing allows users to incrementally refine individual versions after exploring different revision directions. Once a node is confirmed, it turns purple while unconfirmed nodes remain gray, visually distinguishing revision states. Three fine-tuning operations are available: toggling previously applied strategies, providing customized prompts (e.g., "try a different metaphor" or "make this more concise"), and merging two versions to preserve strong elements from each. Visual. These operations support gradual, local refinement within the exploratory space, enabling users to evolve strategy use without committing prematurely.

4.3.2 "Muse" Reflective Feedback (DG2& DG3). (Figure 5(3)) To support DG2 and DG3, the Muse agent monitors user behaviors—such as node confirmations, strategy selections, and engagement–exposition choices—and synthesizes them into structured feedback. This feedback highlights strengths, weaknesses, editing patterns, and strategy suggestions, offering a clear channel for reflection. Users can accept or reject suggestions, and their responses are fed back to the

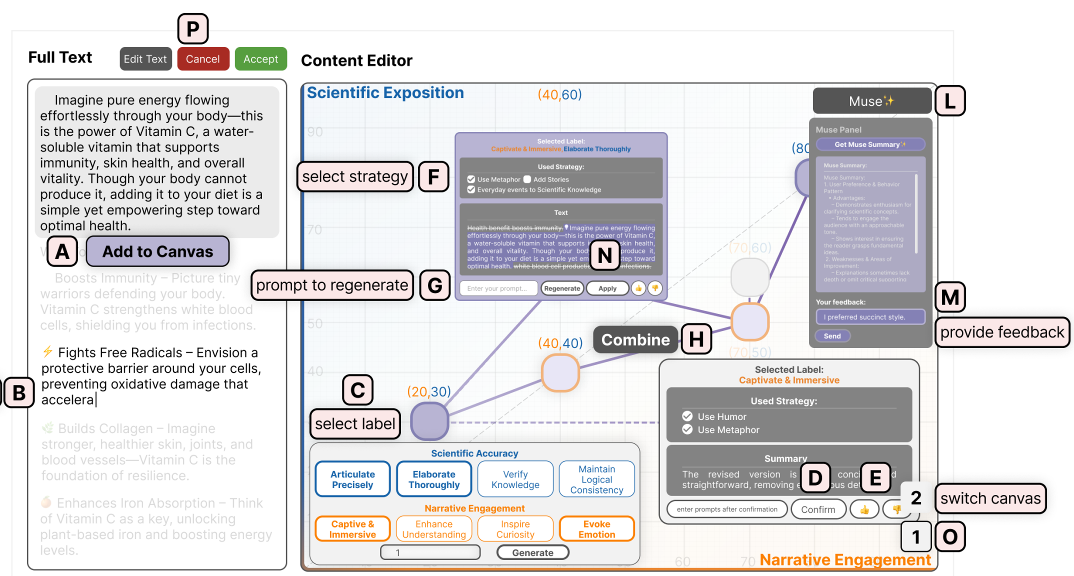

Fig. 4. The SpatialBalancing interface has two main sections: a text editor on the left for placing and directly editing source text (B), and a canvas on the right for revising selected segments (A). In the center, a visualization tracks iteration scores across narrative engagement and scientific exposition for multiple LLM-generated versions. Once a segment is confirmed for revision, users assign labels (C) that guide editing directions and generate revision nodes. Within each node, content can be refined by entering custom prompts (G), switching strategies (F), or combining strategies from different nodes (H). Edits can be applied (N) to update the original text and view the full article. Muse (L), in the canvas's top-right corner, provides an overview of revision history and accepts user feedback (M), which informs future strategy recommendations. Editing other article sections opens a new canvas; users can switch between revision records via the control in the bottom-right corner (O).

Recommender Agent to refine future recommendations. Muse functions as a reflective layer over the exploratory space, supporting trajectory-aware reflection without prescribing edits.

4.4 Backend and Implementation

The backend of SpatialBalancing comprises several LLM-based agents organized into two main modules: a generation module and a reinforcement module. The overall pipeline is in Figure 6.

4.4.1 Generation Module. This module begins by capturing the user's context and their selected modification labels. The system then proceeds into iterative processing handled by the following agents:

Recommender Agent: The recommender agent's core function is to generate multiple strategy combinations based on a user-selected label. When a user chooses a label, the agent analyzes the current textual features to identify the

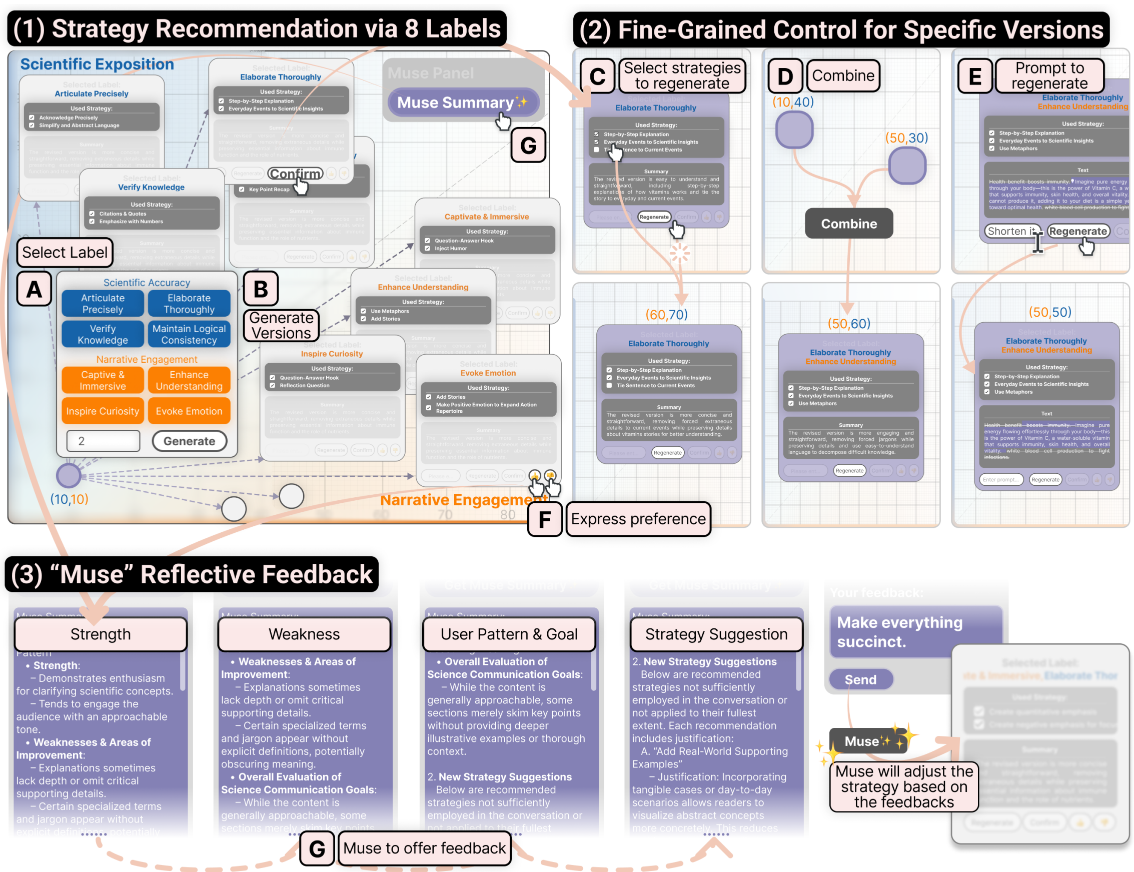

Fig. 5. (1) Strategy Recommendation via Eight Labels: SpatialBalancing offers eight revision labels—four enhancing narrative engagement and four strengthening scientific exposition. users can select one or more labels and specify the number of versions to generate under each; (2) Fine-Grained Control: Generated nodes can be refined by adjusting the applied strategies, merging nodes to combine labels, or entering custom prompts for tailored edits; (3) "Muse" Reflective Feedback: Muse provides iterative feedback on strengths, weaknesses, user patterns and goals, and strategy suggestions. users can endorse or reject this feedback, enabling the system to adapt future recommendations to their preferences.

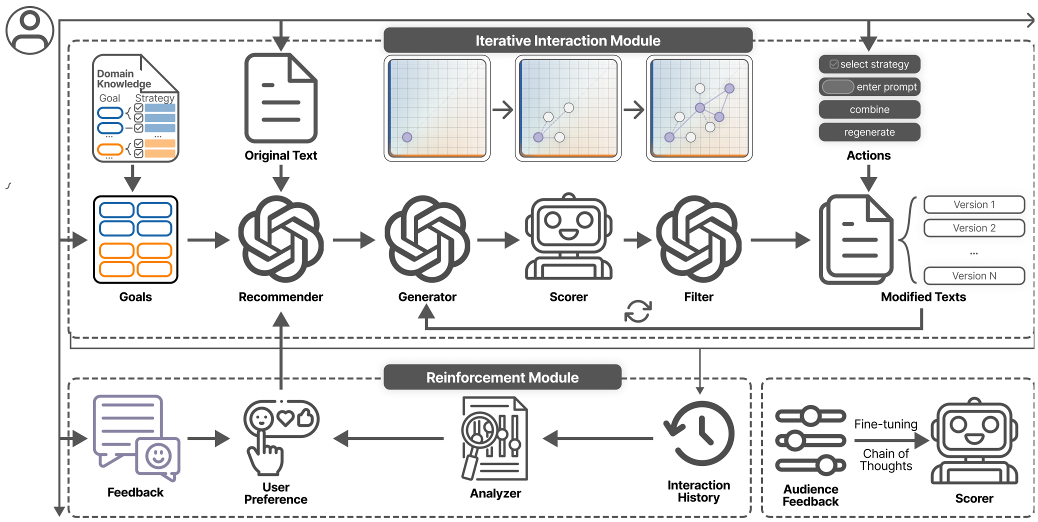

Fig. 6. SpatialBalancing backend overview. SpatialBalancing consists of two core modules: (1) The Iterative Interaction Module, where LLM-based agents—Recommender, Generator, Scorer, and Filter—collaboratively produce and evaluate multiple content versions based on narrative engagement and scientific exposition; and (2) the Reinforcement Module, which captures user feedback and inference based on interaction history of user behaviors to refine strategy recommendations through the Analyzer agent. This architecture supports adaptive text revision.

best combination from its associated strategy set (Section 3.1). Prompts are constructed using in-context learning and chain-of-thought principles based on the strategy design space (Table 5). The agent considers several factors when recommending strategies for each label, including strategy definitions, usage guides, examples, and the original text's role within the broader context of the entire text to recommend the most suitable strategies. The final output consists of multiple strategy combinations, which are then passed to the scorer to filter and select the top-scoring versions that has higher scientific exposition or narrative engagement score.

Generator Agent: The generator agent create child nodes based on user input instructions. When generating new content, the generator receives two types of input to form a new node: (1) strategy recommendations from the Recommender Agent, which are used to guide the generation of revised text that aligns with the user's chosen direction (Labels). The generator adopts in-context learning, referencing the recommended strategies' definitions, usage guidelines, and examples to perform content modifications based on the previous node (adopted from Section 3.1); and (2) user-specific refinements passed from the front end during regeneration. These refinements may include prompt adjustments, combining nodes, or deactivating particular strategies.

Scorer Agent: The scorer simulates real-time audience feedback by evaluating each generated version along two axes: Narrative Engagement (X) and Scientific Exposition (Y).

To support this, we curated a high-quality dataset of 45 science texts from five common science communication domain, varying in length and narrative style. Each text was revised by a science communication expert and annotated by 27 participants who perform as the audience using a rubric developed by three domain experts. The rubric incorporated sub-dimensions of narrative engagement and scientific exposition. Scores were normalized to a 0–100 scale and used to fine-tune a GPT-4o model via a small-sample learning strategy[1]. This enables the scorer agent to give score to resemble

[1]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

human audience across both scientific exposition and narrative engagement. The scorer agent is powered by this fine-tuned GPT-4o model. Details on dataset construction and model training are provided in Appendix A.2.

As such, we acknowledge that the scorer, trained on a small curated dataset. The scoring feedback should be interpreted as an indicative signal for interface interaction and decision-making support, rather than an objective or universal evaluation of the quality of science communication.

To validate the reliability of the scoring mechanism, we conducted a technical evaluation comparing the accuracy of fine-tuned and non-fine-tuned scorers in simulating audience ratings. As shown in Table 2, the fine-tuned scorer exhibited much higher agreement with human ratings (r=0.90/0.91, RMSE≈6–7) than the non-fine-tuned model (r=0.84/0.57, RMSE=22–31). Detailed evaluation detail is provided in Appendix A.2.

Table 2. Evaluation of the similarity between fine-tuned and original GPT-4o models' scores and human scores.

Model	**Pearson Correlation**		**RMSE**	
	Engagement	Exposition	Engagement	Exposition
w/ FT	**0.90**	**0.91**	**6.48**	**7.02**
w/o FT	0.84	0.57	22.48	30.90

Filter Agent: This agent uses the scorer's outputs to select the top-k versions that best meet the user's expectations. Filter Agent ensures that the selected outputs not only fulfill the intended modification chosen direction (Labels) and achieve high scores but also filter out generated failures and low-quality content. This prevents content redundancy and enhances overall generation quality.

4.4.2 Reinforcement Module. Since user iterations form a tree of nodes enriched with valuable data (selected labels, prompts, likes /dislikes, and feedback), we developed an analyzer agent to harness both the explicit and implicit signals from these interactions. The analyzer agent captures behavioral data during the iterative process and uses chain-of-thought prompts to interpret user revision behavior.

Analyzer Agent: The analysis pursues two main goals: (1) identifying common editing patterns, including stylistic preferences, trade-offs between scientific exposition and narrative engagement, and individual user strengths or weaknesses; and (2) uncovering alternative or underused strategy directions. These insights are passed to the Muse component (Section 4.3.2). After the user provides feedback on the LLM's suggestions through Muse, the Analyzer Agent incorporates this real-time feedback (e.g., approvals or further edits) and updates the Recommender Agent accordingly. This process refines subsequent strategy recommendations, ensuring that each iteration aligns more closely with the user's preferences and habits. The feedback loop enables the system to adapt continuously to personal writing habits while balancing narrative engagement and scientific exposition throughout the revision process.

4.4.3 Implementation. SpatialBalancing is implemented as a web application, with a Python-based backend developed using Flask[2] framework and a frontend built using ReactFlow[3].

For the AI agents, we employ different LLMs tailored to their functional roles. The recommender, generator, and filter agents are powered by the GPT-4o-mini model, optimized for fast, high-quality content generation. The analyzer agent, which requires deeper reasoning to interpret user behavior and editing patterns, is supported by the GPT-o1 model—a

[2]https://flask.palletsprojects.com/en/stable/
[3]https://github.com/wbkd/react-flow/

reasoning-oriented LLM. For the scorer agent, it is powered by a fine-tuned GPT-4o model using a small-sample learning strategy[4]. The frontend into predefined prompt templates and communicates with the remote LLMs to obtain results. This modular design allows us to tailor agent behavior based on context while maintaining flexibility in prompt construction and LLM selection. The detailed use of prompts in the backend can be found in the Appendix A.7.

5 User Study

To better understand how SpatialBalancing's spatial externalized visualization design reshapes writers' cognition and human–AI collaboration during LLM-assisted science communication, we conducted a controlled user study comparing SpatialBalancing with a baseline LLM-supported editing workflow. Our goal was to examine how spatial externalization features shape how writers reason, reflect, and iterate during revision, and to derive design insights for interfaces that better support complex revision processes in the process of co-creation with AI. This study addresses two research questions:

RQ1: *How do spatial externalization features shape users' cognitive processes during LLM-assisted iterative revision?*

RQ2: *What interaction tensions and user arise from spatial externalized revision interfaces?*

5.1 Participants

Rather than representing professionally trained science communicators, our participants reflect a growing group of experienced but non-expert science communication creators. To support this, we recruited 16 participants (9 male, 7 female; aged 24–31, M = 26.9, SD = 2.0), all of whom held postgraduate degrees or higher. Many participants were PhD students, postdoctoral researchers, or early-career faculty affiliated with a local university. They all have some experience in creating science communication content and are familiar with using LLM in writing. The demographic information of these participants are in Appendix A.5.

5.2 Procedure

Each study session began with a live demonstration of the system. Participants were encouraged to explore the interface, try out features, and ask questions. During this walkthrough, the task objectives were also explained.

Each participant completed four text editing tasks: two using the SpatialBalancing system and two with the baseline. The texts were selected to represent two common styles of science communication: expository (e.g., "How mRNA Vaccines Work," "Criteria for Animal Domestication") and narrative storytelling (e.g., "Discovery of Archimedes' Principle," "Living and Thriving with ADHD"). Participants were asked to imagine two specific scenarios: (1) for the expository text: "I have a scientific narrative. How can I make it more engaging and interesting for an online science video?" (2) for the narrative storytelling text: "I have a story as an online science video narrative. How can I link it with more scientific concepts and add scientific credibility?" These two scenarios reflect two common starting points in science communication practice: revising from academically oriented, exposition-heavy scientific content, and developing science narratives from everyday experiences or popular media contexts [18].

The length of each text averaged 297.75 words (SD = 19.64). The complete versions of the source texts used for the editing tasks are provided in Appendix A.3. To ensure balanced exposure and mitigate order effects or personal topic preferences, we counterbalanced both the system order (SpatialBalancing vs. baseline) and the text type assigned to each system. Thus, each participant edited one expository and one narrative text under each system condition.

[4]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

Throughout the tasks, participants were encouraged to think aloud, verbalizing their thoughts, reasoning, and feelings as they interacted with the systems. All sessions were screen-recorded, and system interaction logs—such as button clicks (e.g., label selections, generate, regenerate, prompt input, combine)—were automatically captured for the SpatialBalancing condition.

The baseline system used in this study was an interface consisting of a text editor and a conversational agent (powered by GPT-4o) that supported inline editing and suggestions from LLM. In both conditions, participants were provided with an Excel file containing a comprehensive strategy table. This table included the strategy name, definition, usage instructions, examples, and corresponding labels. Participants were encouraged to use this table as a reference and to copy-paste content into the prompt area as needed during the tasks. As such, the baseline served as a conservative comparison, allowing us to examine how making goals, strategies, and revision states explicit and externalized changes users' cognitive processes and collaboration patterns with LLMs.

5.3 Post-Task Survey and Instruments

After completing both conditions, participants completed a post-task survey with standardized instruments: the System Usability Scale (SUS) [6], NASA-TLX for workload [29], and the Creative Self-Efficacy Index (CSI) [10], with one item adapted to: "I think this system supported me in developing ideas or text collaboratively." We also ask participants to evaluate the usefulness of the main design features of SpatialBalancing using eight questions.

Besides, we developed a concise co-creation survey targeting two metacognitive constructs from cognitive psychology [22, 61]. Metacognitive knowledge assessed awareness of cognitive goals (e.g., "I am aware of my writing goals during the editing process"). Metacognitive regulation captured planning, monitoring, and evaluation [54] (e.g., "I set specific goals for the narrative," "I reflect on editing strategies while using the AI tool," and "I reviewed the narrative to assess how well it communicated scientific content"). These items were adapted from the Metacognitive Awareness Inventory [61] and aligned with recent insights into AI-induced metacognitive demands. To measure perceived control during co-creation, we included items inspired by Human-AI interaction principles [69], focusing on participants' influence over outputs and narrative direction. Perceived autonomy was assessed according to Self-Determination Theory [16], addressing decision-making freedom, expressive latitude, and resistance to system pressure. The full list of items on metacognition, perception of control and autonomy is provided in Appendix A.4.

All instruments (NASA-TLX, SUS, CSI, and co-creation survey) employed a 7-point Likert scale. After task completion, each participant joined a 15-minute semi-structured interview designed to capture deeper insights into cognitive processes, feature usage, perceived system value, and moments of difficulty or breakthrough. These interviews complemented survey responses and enriched our understanding of user experience across both conditions.

6 Results

6.1 RQ1: How do spatial externalization features shape users' cognitive processes during LLM-assisted iterative revision?

Drawing on Kirsh's theory of *thinking with external representations* [37], we analyze how SpatialBalancing's features of spatial externalization reshape users' cognitive processes during LLM-assisted revision. As summarized in Table 3, these features transformed iterative revision from an internally managed, reactive process into a spatially navigable activity that supported goal orientation, trajectory-based metacognitive control, and low-cost exploration.

Type of Spatial Externalization	Spatial Externalization Feature	Cognitive Function (Kirsh [37])	Observed Reasoning and Behavior	Representative Evidence
Rhetorical Goals	**2D coordinate space** (scientific exposition × narrative engagement)	*Persistent referents; re-representation of abstract goals*	Externalized rhetorical trade-offs as a stable design state that users could continuously reference, helping them remain oriented to competing goals and avoid drifting into single-direction revisions	"The coordinate graph keeps me from getting lost balancing the two dimensions during revisions" (P3); "I refer to the scores to decide which dimension I need to improve—otherwise I might just keep revising in one direction without noticing as I do in baseline" (P12)
	Strategy labels aligned with axes	*Explicit encoding of strategies; action scaffolding*	Made rhetorical goals actionable by mapping abstract intentions to concrete revision moves; helped users recognize available strategies and reduced the effort of deciding how to revise	"The labels make me realize what kinds of things I should be doing instead of getting lost in details" (P1); "It gave me methods I hadn't considered before" (P12); "The strategies are packaged—I just click and go" (P7)
Iterative Revision Trajectories	**Node-based version layout with visible scores**	*Reduced inferential cost; calibration through comparison*	Enabled side-by-side comparison across multiple versions; scores functioned as indicative reference points to support judgment and prioritization rather than optimization toward a single metric	"I can see strengths and weaknesses by comparing the score of different nodes, not just reading one version" (P8); "Now I first check whether the engagement score is higher compared with previous nodes before reading carefully" (P10); "Coordinate scores help me align edits with my standards and visually track progress. Seeing engagement scores rise reinforces my decisions and makes me feel that I am heading in the right direction." (P3)
	Persistent revision traces with spatial movement	*Trajectory-based reasoning; lowering control cost*	Supported reflection across iterations by making revision history visible as a trajectory of movement, allowing users to interpret progress, regression, and compensatory adjustments between goals	"I can see where each step leads and go back to earlier versions" (P2); "Each version becomes a reference point rather than something I have to remember" (P13); Iterative shifts across axes observed in Fig. 7
Exploratory Space	**Spatial Parallel Prototyping workspace**	*Changing cost structure of exploration*	Lowered the cost of experimentation by enabling non-linear branching, parallel exploration, and reversible decisions without committing to a single path	Higher CSI Exploration and Enjoyment scores; "It gave me room to play and test different directions with low cost" (P11); "I can try several versions of editing direction and still come back to earlier ones to make editing in another direction" (P6); "By selecting different labels, I can explore multiple revision directions, while adjusting strategies or prompts to personalize the edits. This gives me a strong sense of creative flexibility (P1)."

Table 3. How system interface design supports thinking with external representations [37], linking spatial externalization features to cognitive functions and observed reasoning behaviors in LLM-assisted revision.

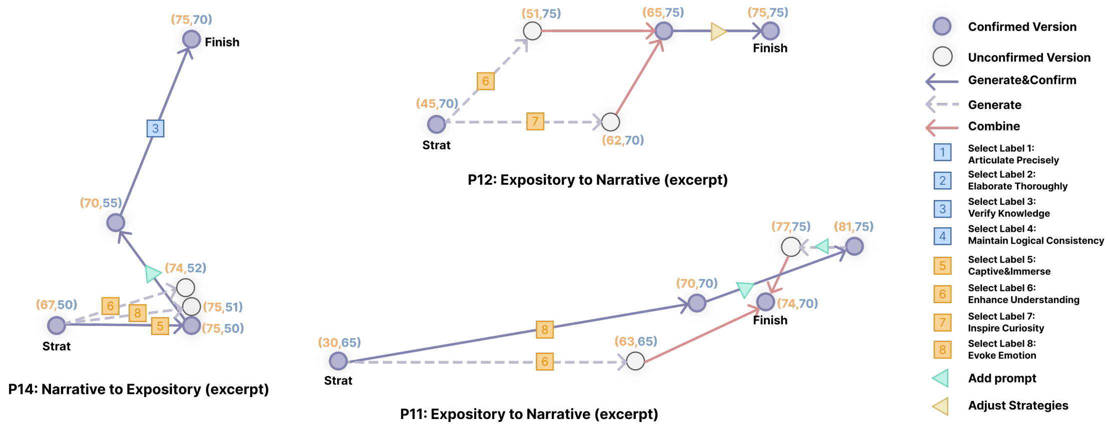

Fig. 7. Visualization examples of segment revisions from P11, P12, and P14.

6.1.1 2D Spatial Externalized Visualization Supports Orientation to Rhetorical Goals. Externalizing scientific exposition and narrative engagement as persistent visual dimensions helped participants remain oriented to competing rhetorical goals throughout revision. Rather than reasoning about balance implicitly or retrospectively, participants treated the coordinate space as a stable reference state that made trade-offs continuously visible (Table 3, Row 1). This reduced goal drift commonly observed in prompt-only workflows and supported focused prioritization during editing(P1, P8, P12).

Strategy labels further operationalized these goals by encoding abstract intentions into actionable revision moves, helping users, especially less experienced writers decide how to revise rather than just whether to revise (P1, P4, P7, P12) (Table 3, Row 2). Among all evaluated features, the two-axis feedback($M = 5.94$, $SD=1.18$) and the strategy labels ($M = 5.81$, $SD=1.17$) were perceived as the most useful, receiving the highest mean ratings with relatively low variance, highlighting their central role in supporting users' revision decisions (Appendix Figure 12).

6.1.2 Spatial Externalized Visualizing Revision Trajectory Enables Metacognitive Control and Confidence Across Iterations. Beyond moment-to-moment orientation, spatial externalization supported metacognitive control across iterations by visualizing the revision trajectory through externalizing the available choices and decisions. Quantitatively, participants using SpatialBalancing reported significantly higher levels of metacognition in reflecting on their own strategies and adjusting strategies during editing (Q3, Q4; see Table 8).

Participants framed revision as a trajectory-based process, deliberately advancing toward one rhetorical goal and then compensating toward the other to restore balance (as shown in Appendix A.6 Figure 11 and Figure 7). This behavior reflects metacognitive control, as writers monitored the effects of prior edits and adjusted subsequent strategies accordingly. Rather than treating generations as isolated outputs, they used externalized cues to track revision states over time and coordinate strategy shifts across iterations, enabling reflective, goal-directed revision.

Qualitatively, the node-based version layout with visible scores and the persistent revision traces with spatial movement jointly supported decision making and process-level control during iterative revision (Table 3, Rows 3–4). By enabling side-by-side comparison across versions, visible scores reduced inferential cost and provided indicative reference points that helped participants judge relative strengths, prioritize revision directions, and decide where to invest attention (P1, P8, P10, P3). Beyond local decisions, persistent spatial traces externalized revision history as a trajectory, allowing participants to interpret progress, regression, and compensatory shifts between rhetorical goals (P2, P13).

Furthermore, scores were used for calibration rather than optimization, reinforcing confidence in the revision process. Just as P3 mentioned, "coordinate scores help me align edits with my standards and visually track progress. Seeing Engagement scores rise reinforces my decisions and making me feel that I am heading in the right direction." By making progress perceptible across iterations, externalization reduces epistemic uncertainty about whether local edits contribute to higher-level goals, thus making participants feel more confident.

6.1.3 Spatial Externalized Exploratory Space Changes the Cost Structure of Exploration. Externalizing the exploratory space also supports creativity. Participants rated SpatialBalancing significantly higher in Exploration and Enjoyment on the CSI questionnaire (Figure 9), without increases in perceived cognitive load (NASA-TLX; Table 4). Qualitative insights indicate that the shared spatial workspace enabled non-linear branching, parallel comparison, and reversible decisions, lowering the risk and effort associated with experimentation (Table 3, Row 5). The free exploratory space also allowed users to explore multiple revision directions simultaneously, encouraging playful testing and occasional conceptual shifts that would be less likely in linear prompt–response workflows.

		SpatialBalancing		Baseline		Statistics	
		mean	std	mean	std	p-value	Sig.
NASA-TLX [29]	Mental Demand	4.63	1.36	4.19	1.68	.404	—
	Physical Demand	3.19	1.60	2.63	0.96	.261	—
	Temporal Demand	2.63	1.36	3.19	1.38	.343	—
	Effort	3.94	1.39	4.44	1.79	.241	—
	Performance	5.13	0.89	4.88	0.96	.372	—
	Frustration	2.88	1.59	3.00	1.32	.724	—
SUS [6]	Q1: use frequently	5.13	1.54	4.38	1.36	.155	—
	Q2: unnecessarily complex	3.00	1.41	2.94	0.85	.899	—
	Q3: easy to use	4.94	1.69	4.88	1.15	.964	—
	Q4: need support	3.94	1.91	2.81	1.87	.031	*
	Q5: function well integrated	5.13	1.26	3.44	1.36	.003	**
	Q6: inconsistency	3.06	1.39	3.25	1.53	.719	—
	Q7: learn to use quickly	4.88	1.59	5.06	1.44	.604	—
	Q8: awkward	2.44	1.26	2.50	1.37	.927	—
	Q9: confident	4.50	1.32	4.50	1.37	.812	—
	Q10: need learning	3.81	1.56	3.38	1.89	.397	—
	Overall Score	70.78	29.70	68.44	26.94	.729	—

Table 4. The statistical results of NASA-TLX and SUS questionnaires. ($*$: $p < 0.05$ and $**$: $p < 0.01$).

6.2 RQ2: What interaction tensions and user expectations arise from spatial externalized revision interfaces?

6.2.1 Balancing Externalized Guidance and User Judgment. Participants described how the system's visual and scoring feedback may influence their evaluation practices in subtle ways. While the coordinate axis enabled intuitive comparisons between revisions, some participants noted that the visibility and immediacy of scores could reduce their depth of textual engagement. As P4 reflected, " When using the system, I outsourced a large part of the thinking process to the AI. It's faster and more efficient, but I also tend to think less carefully about the output as I trust the score results

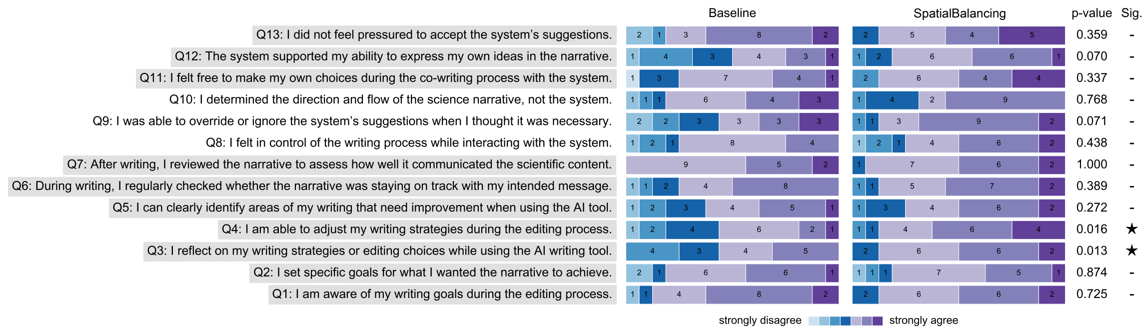

Fig. 8. Results of the Metacognition (Q1–Q7), Control (Q8–Q10), and Autonomy (Q11–Q13) questionnaires (p < .05 marked with *; p < .01 with **). Significant differences were observed in Metacognition: RQ3 (M = 5.50 (SpatialBalancing) vs. 4.63 (Baseline), p = .013) and RQ4 (M = 5.69 vs. 4.56, p = .016); marginal differences in Control: RQ9 (M = 5.63 vs. 4.75, p = .071) and Autonomy: RQ12 (M = 5.25 vs. 4.44, p = .070).

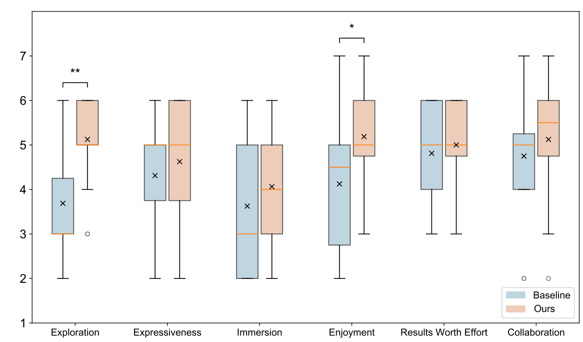

Fig. 9. The results of CSI questionnaire. ($*$: $p < 0.05$ and $**$: $p < 0.01$). Participants rated SpatialBalancing significantly higher in terms of "Exploration" (M = 5.13 (SpatialBalancing) vs. 3.69 (Baseline), p = .004) and "Enjoyment" (M = 5.19 vs. 4.13, p = .039)

more than I did with the baseline. In baseline, I would read text more carefully and make judgments by myself." This suggests that while externalized scoring streamlines comparison, it can also shift evaluative effort away from close reading toward greater reliance on system-provided judgments.

Others expressed a degree of caution about over-relying on the scores. P16 noted that while the visual feedback was useful, "the scores are indicative rather than definitive. They sometimes do not reflect the actual quality of the generation and still require human judgment. " P7 also noted that although the coordinate view provides scores, they still read the text carefully and reconcile the system's feedback with their own standards. As a result, they sometimes chose versions located at intermediate positions rather than pursuing extreme scores. These reflections suggest a potential tension: while the system offers accessible and actionable feedback, its effectiveness depends on users' ability to critically interpret the signals rather than accept them at face value.

Concerns about the interpretability of scoring were also raised. As P14 said, "Sometimes I don't know what an increase in score actually means. I can't tell whether each label contributes differently to the score or what specific content led to a higher score. I want to understand the logic behind the numbers. " This suggests the interpretability of the scores and the changes made to them also needs improvement.

6.2.2 *Seek More Flexible and Adaptivity of Externalization in Use.*

While the eight-label set was seen as a helpful starting point, more experienced participants felt it could be more flexible to be customized to better support their advanced needs. P3 shared that: P1, P3, P2, and P14 wished they could combine or tailor underlying strategies to form customized labels to align more closely with their specific intentions.P1 expressed a desire to curate combinations of strategies based on their own habits, and to flexibly create new combinations to support more personalized needs. P14 also noted, "In addition to the current style-focused labels, it would be helpful to include others that target areas in writing revision like grammar or tone." Together, these responses reflect a tension between predefined externalized guidance and users' desire for greater agency.

Participants also reported that repeated exposure to the coordinate scores helped them develop a personal reference range, allowing them to recognize patterns in their own writing habits over time (P3, P4). Rather than treating the scores as absolute targets, they used them to understand where their typical writing tended to fall and how revisions shifted that position. As P3 suggests, "if the visualization can provide further visual indication of the score ranges preferred by specific reader groups, it could enable more informed adjustments by helping authors intentionally move the revision points toward positions that better align with different audience expectations. This indicates that participants appropriated the coordinate scores as a personalized, evolving reference system rather than fixed evaluative benchmarks, using repeated exposure to calibrate their own writing tendencies and reason about audience-specific adjustments.

6.2.3 *More Proactive and Grounded Feedback in the Revision Process.*

While participants appreciated what Muse could already do to help reflect on the whole revision process (P2, P6, P13, P14), P2 wanted more real-time dialogue: "I wish it were more interactive—like chatting with someone who helps me reflect as I go during the revision process." P13 and P14 also expected the system to proactively offer assistance, even before they explicitly recognized the need for help. The log data further indicates that participants tended to use the Muse function primarily at the final stage of their revision and only once in most cases (Appendix Figure A.6). This points to the need for more proactive and embedded reflective interactions rather than relying on users to initiate reflection themselves.

Participants also wished that Muse could give more personalized and context-specific feedback in the revision process (P1, P8). "Right now, Muse gives high-level suggestions, " P8 said. "But it'd be more useful if it could point to which step or decision was strong or weak, and explain why. " This suggests that participants seek feedback that is grounded in specific revision actions and their underlying rationale.

7 Discussion

7.1 Design Insights for Externalization in Human–AI Writing Interfaces

7.1.1 *Design Insight 1: Mitigating Metacognitive Laziness of Relying on Externalization.*

According to distributed cognition, reflection and problem solving are not confined to an individual's internal reasoning, but are distributed across interactions among users, AI models, and external tools [28]. In such distributed cognitive systems, merely providing access to knowledge or suggestions is insufficient. Users must also be able to understand, monitor, and regulate how this knowledge is produced, interpreted, and applied within the system—an ability that lies at the core of metacognition [64]. Our user study suggests that externalizing rhetorical goals and revision trajectories can effectively enhance metacognitive regulation during LLM-assisted revision. By making abstract goals and revision progress perceptible through spatial cues, users were better able to reflect on their editing strategies, adjust revision directions across iterations, and maintain a sense of process-level control. These findings align with prior work showing that external representations can support planning, monitoring, and evaluation by reducing the inferential burden of tracking change internally [37].

However, the same externalization properties also reveal a potential tension. Highly legible and actionable feedback—such as explicit scores and spatial comparisons can simultaneously support reflection and displace reflective effort. Several participants reported that they began to rely more on system-provided cues to make decisions, sometimes at the expense of close reading and independent evaluation of the text. In these moments, evaluative judgment shifted from users' own critical reasoning toward system-generated signals. This echoes prior findings that frequent reliance on LLM feedback may encourage over-trust and lead to "metacognitive laziness," in which users reduce self-regulation and critical engagement with the task [20, 64]

Together, these findings highlight an important design challenge that rather than treating cognitive offloading as an unqualified benefit, designers should carefully consider what aspects of cognition are externalized and how users are invited to engage with them [64]. Design consideration using this kind of externalized visualization features can be made: (1) Designing feedback as reflective prompts rather than prescriptive, for example, by framing scores as indicative signals that invite interpretation, comparison, or questioning, instead of optimization targets; (2) Supporting moments of deliberate re-engagement, such as encouraging users to articulate why they accept, reject, or override system suggestions, thereby reinforcing evaluative ownership; (3) Providing adjustable levels of guidance, allowing users to control when and how much evaluative feedback is visible, so that reliance on external cues can be modulated over time and expertise levels. (4) Making the basis of system feedback more interpretable, helping users understand why certain revisions shift scores, which can transform externalized metrics from authority signals into learning resources.

7.1.2 Design Insight 2: Preserving Agency through Adaptive Mixed-Initiative Externalization. Scaffolding through externalization is effective for supporting rapid prototyping and reducing decision overhead in LLM-assisted writing, particularly in early stages of revision. By packaging strategies into higher-level labels, the system helps users quickly explore and compare revision directions. However, as users gain experience, fixed scaffolds can become constraining, no longer aligning with their evolving intentions, personal writing habits, or situational goals. Our findings show that experienced users wanted to move beyond predefined labels by curating and recombining underlying strategies, treating externalized structures not as fixed guidance but as resources to be reshaped.

These findings suggest that externalization should function as a flexible and adaptive substrate, rather than a static scaffold. Interfaces should support user-driven customization of externalized elements (e.g., allowing users to create or curate personalized labels), while also enabling system-driven adaptation based on observed interaction patterns. For example, by reflecting stable writing patterns back into the visualization—such as indicating personal reference zones or audience-specific target regions within the exploratory space—the system can adapt externalized cues to users' evolving goals and habits. In this way, externalization shifts from prescribing ideal targets to supporting situated self-regulation. Previous work suggests that agency in human–AI co-creation fluctuates across the creative process [56], so designs that adapt externalization in this manner preserve the efficiency benefits of scaffolding while gradually restoring user agency, enabling more individualized, reflective, and sustainable writing practices in human–AI collaboration.

7.1.3 Design Insight 3: Providing Proactive In-Situ Reflective Support. While reflective support is essential for helping writers make sense of iterative revisions, relying on users to explicitly initiate reflection can limit its effectiveness. In our results, we found that when reflection relied primarily on user initiative, participants were less likely to engage in it proactively. Instead, they expressed a preference for more step-by-step, in-situ reflective support integrated into the revision process.

This finding points to the need for proactive mechanisms that surface reflective support at appropriate moments within the revision process. For example, rather than relying on users to pause and reflect on their own, systems should proactively trigger reflection at natural breakpoints in revision, such as when users compare alternatives, confirm a revision, or shift revision direction, instead of expecting users to stop. In addition, reflective support should be grounded in visible artifacts of revision [75, 76] such as generated alternatives, score changes, or spatial movements—to make reflection concrete and interpretable. For example, when users repeatedly explore multiple versions of a passage without reaching a satisfactory outcome, the system can proactively surface reflective questions besides the revision node to help clarify underlying intentions.

7.2 Limitation and Future Work

We describe several limitations in the study to define the scope of our findings clearly and motivate future work.

7.2.1 Lack of Evaluation on Text Quality and Communication Effectiveness. One limitation of the current study is the absence of a systematic evaluation of the generated texts, though our main focus is the design and deployment of the system. While the system produces revised versions of scientific narratives, we did not assess whether these revisions lead to improvements in quality for science communication purposes. Future studies could investigate whether the generated texts are more engaging and whether they facilitate better knowledge retention among audiences. Objective and subjective measures, such as expert evaluation, audience feedback through deployment, and comprehension tests, could be employed to evaluate the effectiveness of the texts in real-world science communication settings to further validate the effectiveness of the system for the output of texts.

7.2.2 Evaluation Dependency on Proxy Scores. To demonstrate SpatialBalancing with minimal evaluation overhead, we adopted a low-cost approach that uses model-generated proxy scores to approximate audience feedback on scientific exposition and narrative engagement. These proxies were intended to support comparative reasoning and iterative decision making during revision, rather than to represent comprehensive or definitive audience judgments. While such scores may reflect the expectations of a particular participant group, they cannot capture the full diversity of real-world audiences or contexts (e.g., classroom learning vs. online videos). Accordingly, the current scoring mechanism should be understood as a design probe, and future work should validate and extend it with audience- and context-specific evaluation methods.

7.2.3 Methodological Limitations. This work has common methodological limitations including the short-term nature of system testing which may not reveal long-term adoption patterns, and the relatively homogeneous participant demographics that may not represent all potential user groups. Future work will aim to address the previously mentioned and these limitations through more comprehensive evaluations.

8 Conclusion

Our results show that spatial externalization reshape how writers reason about LLM-assisted revision. By externalizing rhetorical goals and revision history in an exploratory spatial workspace, participants treated revision as a trajectory rather than a series of isolated edits, enabling sustained goal orientation, metacognitive control across iterations, and low-cost exploration of alternatives. The two-dimensional feedback functioned as navigational cues—supporting calibration and reflection, rather than prescriptive optimization signals. At the same time, participants surfaced tensions around over-reliance on externalized scores and the need for more flexible, adaptive forms of externalization. Together, these findings suggest that the value of spatial exploratory interfaces lies not in generating better revisions per se, but in supporting writers' ability to navigate, reflect on, and steer complex revision processes over time.

References

[1] Isabelle Augenstein. 2021. Determining the Credibility of Science Communication. arXiv:2105.14473 [cs.DL] https://arxiv.org/abs/2105.14473

[2] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5327–5344. doi:10.18653/v1/2020.emnlp-main.429

[3] Erik Blair. 2015. A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences* 6, 1 (2015), 14–29.

[4] Besma Boubertakh. 2015. Towards Further Experimental Reproducibility: Making A Balance Between Conciseness, Precision and Comprehensiveness in Scientific Communication. *Journal of Neurology and Stroke* 3, 1 (Oct. 2015). doi:10.15406/JNSK.2015.03.00077

[5] Peter Broks. 2006. *Understanding popular science*. McGraw-Hill Education (UK).

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Terry W Burns, D John O'Connor, and Susan M Stocklmayer. 2003. Science communication: a contemporary definition. *Public understanding of science* 12, 2 (2003), 183–202. doi:10.1177/09636625030122004

[8] Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology* 12, 4 (2009), 321–347. doi:10.1080/15213260903287259

[9] Rocco Caferra, Giuseppe Di Liddo, Andrea Morone, and David Stadelmann. 2025. The media morphosis of science communication during crises. *Scientific Reports* 15, 1 (2025), 5506. doi:10.1038/s41598-025-88973-7

[10] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25. doi:10.1145/2617588

[11] John Joon Young Chung and Max Kreminski. 2024. Patchview: LLM-powered Worldbuilding with Generative Dust and Magnet Visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 77, 19 pages. doi:10.1145/3654777.3676352

[12] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Toyteller: AI-powered Visual Storytelling Through Toy-Playing with Character Symbols. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 331, 23 pages. doi:10.1145/3706598.3713435

[13] Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences* 111, Supplement 4 (2014), 13614–13620. doi:10.1073/pnas.1320645111

[14] Michael F. Dahlstrom and Dietram A. Scheufele. 2018. (Escaping) the paradox of scientific storytelling. *PLOS Biology* 16, 10 (10 2018), 1–4. doi:10.1371/journal.pbio.2006720

[15] Andreas W Daum. 2009. Varieties of popular science and the transformations of public knowledge: some historical reflections. *Isis* 100, 2 (2009), 319–332. doi:10.1086/599550

[16] Edward L. Deci and Richard M. Ryan. 2012. Self-Determination Theory. In *Handbook of Theories of Social Psychology*, Paul A. M. Van Lange, Arie W. Kruglanski, and E. Tory Higgins (Eds.). Vol. 1. SAGE Publications, Thousand Oaks, CA, USA, 416–436. doi:10.4135/9781446249215.n21

[17] Anne DiPardo. 1990. Narrative knowers, expository knowledge: Discourse as a dialectic. *Written communication* 7, 1 (1990), 59–95. doi:10.1177/0741088390007001003

[18] Julie S. Downs. 2014. Prescriptive scientific narratives for communicating usable science. *Proceedings of the National Academy of Sciences of the United States of America* 111, Suppl 4 (2014), 13627–13633. doi:10.1073/pnas.1317502111

[19] Lee Ellis. 2022. Improving Scientific Communication by Altering Citation and Referencing Methods. *Journal of Social Science Studies* 9, 1 (2022), 1–1. doi:10.5296/jsss.v9i1.19548

[20] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology* 56, 2 (2025), 489–530. arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13544 doi:10.1111/bjet.13544

[21] Wiebke Finkler and Bienvenido León-Anguiano. 2019. The power of storytelling and video: a visual rhetoric for science communication. (2019). doi:10.22323/2.18050202

[22] John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10 (1979), 906. doi:10.1037/0003-066X.34.10.906

[23] Linda Flower and John R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32, 4 (1981), 365–387. doi:10.2307/356600

[24] Laura Fogg-Rogers, Ann Grand, and Margarida Sardo. 2015. Beyond dissemination - Science communication as impact. 14, 3 (Sept. 2015). doi:10.22323/2.14030301

[25] Eric L Garland, Barbara Fredrickson, Ann M Kring, David P Johnson, Piper S Meyer, and David L Penn. 2010. Upward spirals of positive emotions counter downward spirals of negativity: Insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical psychology review* 30, 7 (2010), 849–864. doi:10.1016/j.cpr.2010.03.002

[26] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009). doi:10.1515/COMM.2009.026

[27] Jean Goodwin and Michael F. Dahlstrom. 2014. Communication strategies for earning trust in climate change debates. *WIREs Climate Change* 5, 1 (2014), 151–160. arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.262 doi:10.1002/wcc.262

[28] James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.* 7, 2 (June 2000), 174–196. doi:10.1145/353485.353487

[29] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143. doi:10.1016/S0166-4115(08)62386-9

[30] Tianle Huang and Will J Grant. 2020. A good story well told: Storytelling components that impact science video popularity on YouTube. *Frontiers in Communication* 5 (2020), 86. doi:10.3389/fcomm.2020.581349

[31] Oksana Ivchenko and Natalia Grabar. 2022. Impact of the Text Simplification on Understanding. In *Challenges of Trustable AI and Added-Value on Health*. IOS Press, 634–638. doi:10.3233/SHTI220546

[32] Roger A. Pielke Jr. 2007. *The Honest Broker: Making Sense of Science in Policy and Politics.* Cambridge University Press, Cambridge, United Kingdom. doi:10.1017/CBO9780511818110

[33] Klemens Kappel and Sebastian Jon Holmen. 2019. Why science communication, and does it work? A taxonomy of science communication aims and a survey of the empirical evidence. *Frontiers in communication* 4 (2019), 55. doi:10.3389/fcomm.2019.00055

[34] Jagdish Kaur. 2012. Saying it again: enhancing clarity in English as a lingua franca (ELF) talk through self-repetition. *Text & Talk* 32, 5 (Jan. 2012), 593–613. doi:10.1515/TEXT-2012-0028

[35] Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman. 2021. Straight From the Scientist's Mouth—Plain Language Summaries Promote Laypeople's Comprehension and Knowledge Acquisition When Reading About Individual Research Findings in Psychology. *Collabra: Psychology* 7, 1 (02 2021), 18898. arXiv:https://online.ucpress.edu/collabra/article-pdf/7/1/18898/835600/collabra_2021_7_1_18898.pdf doi:10.1525/collabra.18898

[36] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135. doi:10.1145/3563657.3595996

[37] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454. doi:10.1007/s00146-010-0272-8

[38] Laura M König, Marlene S Altenmüller, Julian Fick, Jan Crusius, Oliver Genschow, and Melanie Sauerland. 2024. How to communicate science to the public? Recommendations for effective written communication derived from a systematic review. *Zeitschrift für Psychologie* (2024). https://doi.org/10.1027/2151-2604/a000572

[39] Christoph Kueffer and Brendon M. H. Larson. 2014. Responsible Use of Language in Scientific Writing and Science Communication. *BioScience* 64, 8 (06 2014), 719–724. arXiv:https://academic.oup.com/bioscience/article-pdf/64/8/719/8719054/biu084.pdf doi:10.1093/biosci/biu084

[40] Joe Lambert. 2013. *Digital storytelling: Capturing lives, creating community.* Routledge. doi:10.4324/9780203102329

[41] Martha Larson, Eamonn Newman, and Gareth JF Jones. 2010. Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment. In *Multilingual Information Access Evaluation II. Multimedia Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30-October 2*. 354–368. https://link.springer.com/chapter/10.1007/978-3-642-15751-6_46

[42] Khanh Chi Le, Linghe Wang, Minhwa Lee, Ross Volkov, Luan Tuyen Chau, and Dongyeop Kang. 2025. ScholaWrite: A Dataset of End-to-End Scholarly Writing Process. arXiv:2502.02904 [cs.HC] https://arxiv.org/abs/2502.02904

[43] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. 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 1054, 35 pages. doi:10.1145/3613904.3642697

[44] Robert A. Lehrman and Eric Schnure. 2019. *The Political Speechwriter's Companion: A Guide for Writers and Speakers.* CQ Press, Washington, DC, USA. https://collegepublishing.sagepub.com/products/the-political-speechwriters-companion-2-257507

[45] Norma J. Livo and Sandra A. Rietz. 1986. *Storytelling: Process and Practice.* Libraries Unlimited, Littleton, CO, USA. https://archive.org/details/storytellingproc00livo

[46] Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 782–803. doi:10.1145/3643834.3661587

[47] Tinca Lukan, Bronwen Deacon, and Alina Kontareva. 2024. How TikTok Science Communicators Navigate Norms and Values in the Age of Generative AI. *Elephant in the Lab* (October 3 2024). doi:10.5281/zenodo.13885999

[48] Yao Lyu, He Zhang, Shuo Niu, and Jie Cai. 2024. A Preliminary Exploration of YouTubers' Use of Generative-AI in Content Creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–7. doi:10.1145/3613905.3651057

[49] Roger Maskill. 1988. Logical Language, Natural Strategies and the Teaching of Science. *International Journal of Science Education* 10, 5 (1988), 485–495. doi:10.1080/0950069880100502

[50] Daniel G. McDonald. 2014. Narrative Research in Communication: Key Principles and Issues. *Review of Communication Research* 2, 1 (2014), 115–132. doi:10.12840/issn.2255-4165.2014.02.01.005

[51] Julia Metag, Florian Wintterlin, and Kira Klinger. 2023. Science Communication in the Digital Age-New Actors, Environments, and Practices. *Media and Communication* 11, 1 (2023), 212–216. doi:10.17645/mac.v11i1.6905

[52] Jesús Muñoz Morcillo, Klemens Czurda, and Caroline Trotha. 2016. Typologies of the popular science web video. *Journal of Science Communication* 15 (May 2016), A02. doi:10.22323/2.15040202

[53] National Academies of Sciences, Engineering, and Medicine. 2017. *Communicating Science Effectively: A Research Agenda*. National Academies Press, Washington, DC, USA. doi:10.17226/23674

[54] Chenghai Qin, Ruru Zhang, and Yanling Xiao. 2022. A questionnaire-based validation of metacognitive strategies in writing and their predictive effects on the writing performance of English as foreign language student writers. *Frontiers in Psychology* 13 (2022), 1071907. doi:10.3389/fpsyg.2022.1071907

[55] Fatemeh Rabiee. 2004. Focus-group interview and data analysis. *Proceedings of the nutrition society* 63, 4 (2004), 655–660. doi:10.1079/PNS2004399

[56] Janet F. Rafner, Blanka Zana, Ida Bang Hansen, Simon Ceh, Jacob Sherson, Mathias Benedek, and Izabela Lebuda. 2025. Agency in Human-AI Collaboration for Image Generation and Creative Writing: Preliminary Insights from Think-Aloud Protocols. *Creativity Research Journal* (2025), 1–24. doi:10.1080/10400419.2025.2587803

[57] 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1042, 18 pages. doi:10.1145/3613904.3641899

[58] Marie-Claude Roland. 2009. Quality and integrity in scientific writing: prerequisites for quality in science communication. *Journal of Science Communication* 8, 2 (2009), A04. https://doi.org/10.22323/2.08020204

[59] Gillian Rowe, Jacob B Hirsh, and Adam K Anderson. 2007. Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences* 104, 1 (2007), 383–388. doi:10.1073/pnas.0605198104

[60] Maximilian Roßmann. 2025. Science Correction as a Communication Problem: Insights from Four Theoretical Lenses. *OSF Preprints* (3 February 2025). doi:10.31219/osf.io/82duj_v3

[61] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary educational psychology* 19, 4 (1994), 460–475. doi:10.1006/ceps.1994.1033

[62] Terence A. Shimp. 2010. *Advertising, promotion, and other aspects of integrated marketing communications* (8th ed ed.). South-Western Cengage Learning. https://cir.nii.ac.jp/crid/1970867909787407802

[63] Momin N Siddiqui, Roy D Pea, and Hari Subramonyam. 2025. ScriptShift: A Layered Interface Paradigm for Integrating Content Development and Rhetorical Strategy with LLM Writing Assistants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 532, 19 pages. doi:10.1145/3706598.3714119

[64] Sidra Sidra and Claire Mason. 2025. Generative AI in Human-AI Collaboration: Validation of the Collaborative AI Literacy and Collaborative AI Metacognition Scales for Effective Use. *International Journal of Human–Computer Interaction* 0, 0 (2025), 1–25. arXiv:https://doi.org/10.1080/10447318.2025.2543997 doi:10.1080/10447318.2025.2543997

[65] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. 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 1039, 19 pages. doi:10.1145/3613904.3642754

[66] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. doi:10.1145/3613904.3642400

[67] David H. Torres and Douglas E. Pruim. 2019. Scientific storytelling: A narrative strategy for scientific communicators. *Communication Teacher* 33, 2 (April 2019), 107–111. doi:10.1080/17404622.2017.1400679 Publisher: Routledge _eprint: https://doi.org/10.1080/17404622.2017.1400679.

[68] Gilson Luiz Volpato. 2015. O método lógico para redação científica. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde* 9, 1 (2015). doi:10.29397/reciis.v9i1.932

[69] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300831

[70] Nedra Kline Weinreich. 2010. *Hands-on social marketing: a step-by-step guide to designing change for good*. Sage.

[71] Wikipedia contributors. 2024. Popular science — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Popular_science Accessed: 2025-03-27.

[72] Haijun Xia, Hui Xin Ng, Chen Zhu-Tian, and James Hollan. 2022. Millions and Billions of Views: Understanding Popular Science and Knowledge Communication on Video-Sharing Platforms. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 163–174. doi:10.1145/3491140.3528279

[73] Xian Xu, Leni Yang, David Yip, Mingming Fan, Zheng Wei, and Huamin Qu. 2022. From 'Wow' to 'Why': Guidelines for Creating the Opening of a Data Video with Cinematic Styles. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–20. doi:10.1145/3491102.3501896

[74] Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. 2022. A Design Space for Applying the Freytag's Pyramid Structure to Data Stories. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 922–932. doi:10.1109/TVCG.2021.3114774

[75] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 935, 27 pages. doi:10.1145/3706598.3714316

[76] Chao Zhang, Kexin Ju, Zhuolun Han, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Synthia: Visually Interpreting and Synthesizing Feedback for Writing Revision. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 88, 16 pages. doi:10.1145/3746059.3747703

[77] Yu Zhang, Kexue Fu, and Zhicong Lu. 2025. RevTogether: Supporting Science Story Revision with Multiple AI Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 462, 7 pages. doi:10.1145/3706599.3719888

[78] Yu Zhang, Changyang He, Huanchen Wang, and Zhicong Lu. 2023. Understanding Communication Strategies and Viewer Engagement with Science Knowledge Videos on Bilibili. 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 668, 18 pages. doi:10.1145/3544548.3581476

[79] 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*. ACM, 1–30. doi:10.1145/3586183.3606800

[80] Jelena Šuto, Ana Marušić, and Ivan Buljan. 2023. Linguistic analysis of plain language summaries and corresponding scientific summaries of Cochrane systematic reviews about oncology interventions. *Cancer Medicine* 12, 9 (2023), 10950–10960. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cam4.5825 doi:10.1002/cam4.5825

A Appendix

A.1 Specific Strategies for Science Communication Writing

Table 5. Design Space for Science Communication Writing

Category	Strategy	Definition	Label
Scientific Exposition	(1) Layered Transitions [38, 49, 60, 68]	Use multiple transition words or phrases (e.g., "but," "and," "therefore") within a short span to emphasize logical shifts and contrasts.	4
	(2) Rigorous Source Verification [1, 38, 58]	Cross-check scientific claims and data against reliable, peer-reviewed sources to ensure exposition.	3
	(3) Step-by-Step Explanation [2, 38]	Introduce the core idea first and then progressively add background details, creating a structured learning process.	2, 4
	(4) Acknowledge Uncertainties [32]	Transparently discuss uncertainties, potential biases, or limitations in data and models to build credibility.	1, 2
	(5) Consistent Terminology [39]	Use the same terminology throughout the content to maintain clarity and avoid confusion.	1
	(6) Citations & Quotes [1, 19]	Integrate citations and direct quotes seamlessly to enhance credibility while maintaining narrative flow.	3
	(7) Everyday Events to Scientific Insights [2, 39]	Automatically identify and link theories or knowledge to real-world events or stories mentioned in the text.	2, 3
Narrative Engagement	(8) Question-Answer Hook [21, 30, 40]	Ask a direct question and provide an immediate answer to introduce key concepts clearly and concisely.	5, 6, 7
	(9) Reflection Question [21]	Ask a thought-provoking question that does not require an immediate answer, encouraging reflection and reinforcing key concepts.	5, 7, 8
	(10) Suspense-Driven Reveal [72, 78]	Present a question, problem, or scenario at the beginning and delay its resolution to sustain curiosity.	5, 7
	(11) Use metaphors [21, 39?]	Convey unfamiliar concepts by drawing analogies to more familiar ones.	5, 6
	(12) Inject humor [27]	Use playful language or puns to make the content more engaging and enjoyable.	5, 8
	(13) Add real-world supporting examples [44, 45]	Illustrate abstract concepts using relatable, real-world examples.	5, 6
	(14) Add stories [13, 14, 45]	Use narratives with characters, settings, and plot progression to enhance engagement and memorability.	5, 6, 8
	(15) Add an imagery description [21, 26, 62]	Use vivid, sensory details to help the audience visualize concepts.	5, 6
	(16) Create negative emphasis for focused attention [21, 26, 30, 52]	Highlight extreme negative outcomes to intensify focus and reinforce key lessons.	5, 8
	(17) Make positive emotion to expand action repertoire [21, 25, 26, 52, 59, 70]	Use uplifting messages, particularly in conclusions, to inspire optimism and motivation.	5, 8
Both	(18) Simplify and abstract language [31, 35, 80]	Rephrase complex scientific terminology or detailed descriptions into more general, accessible language without compromising core exposition.	1, 6
	(19) Clarify Key Terms [52, 60]	Define complex or specialized terms at the beginning to establish a shared understanding.	1, 6
	(20) Key Point Recap [21, 52, 67]	Summarize the main points concisely at the conclusion of the content to reinforce memory retention.	1, 4, 6
	(21) Repeat key point(s) or question(s) [4, 34]	Reinforce key concepts by strategically repeating crucial terms or questions.	1, 6
	(22) Emphasize with Numbers [24, 74]	Connect scientific discussions to real-world recent news or trends to enhance relevance and engagement.	1, 2, 3, 8
	(23) Strengthen the Connections Between Content [49, 68]	Ensure smooth transitions between related ideas by using bridging statements or contextual links.	4, 6
	(24) Present Balanced Views [39]	Provide both supporting evidence and counterarguments to present a well-rounded discussion.	2, 6
	(25) Tie Science to Current Events [2, 39]	Connect scientific discussions to real-world recent news or relavant stories.	3, 5, 6

***Lable:** *Scientific Exposition Effects:* 1. Articulate Precisely; 2. Elaborate Thoroughly; 3. Verify Knowledge; 4. Maintain Logical Consistency

Narrative Engagement Effects: 5. Captivate & Immerse; 6. Enhance Understanding; 7. Inspire Curiosity; 8. Evoke Emotion

A.2 Rating Model Construction

Our primary goal in constructing the coordinate axis is to simulate audience feedback so that users can receive real-time evaluations. Therefore, we collected real user feedback on texts with varying characteristics to fine-tune a LLM that can provide scores during the real-time writing process.

Dataset Construction We first built a dataset of popular science texts containing 45 texts (example in section A.2.1) from five commonly seen science communication topics: psychology, economics, geography, history, and physics. For each topic, there are nine texts; three each of long (300 words), medium (150 words), and short (50 words) formats; representing three typical levels of revision granularity in science communication. Within each length category, we included three different levels of narrative transformation: (1) purely expository scientific texts (Expository), (2) fully narrative story-like texts (Story), and (3) an intermediate "infotainment" style (Medium), which is an ideal format in popular science that maintains scientific exposition while incorporating narrative strategies from our design space. All texts were revised by an expert with two years of experience in science communication writing

Score Collection We designed a survey to collect ratings for these texts on two dimensions: Narrative Engagement and Scientific Exposition, two main communication goals in popular science [13]. For Narrative Engagement, we used five subscales: Narrative Presence, Emotional Engagement, Narrative Understanding, Curiosity, and General Narrative Engagement, a survey developed by prior work [8]. For Scientific Exposition, given the lack of mature scales, we measured five dimensions inspired by standards for scientific texts from previous research [13]: Conceptual Clarity, Plausibility, Completeness, and Perceived Factual Correctness. The full questionnaire can be found in the section .

Participants First, we recruited three experts (each with more than one year of experience in creating science narratives) to rate the texts. After rating, they discussed and jointly established a scoring rubric, including benchmarks for each score range from 0 to 10. Next, we recruited 27 participants interested in science communication. We invite experts to establish standards as a reference point for audience ratings, in order to reduce variance in their subjective evaluations of the text. The criteria established by experts are in the Appendix A.2.3.

Survey Results The distribution of scores for the 45 texts is displayed in the Figure 10. It is shown that story-like texts tend to elicit higher narrative engagement but exhibit lower scientific exposition. In contrast, expository texts maintain higher scientific exposition at the expense of engagement. The infotainment style appears to strike a balance between the two. Additionally, longer texts generally perform better in both dimensions, whereas shorter texts show lower overall scores, likely due to limitations in content depth and development.

Final Model Fine-Tuning For each text, we first computed the average score across the five questions within each of the two dimensions and then averaged these scores across all 27 participants. To match the 0–100 scale of the final coordinate axis, the scores were scaled by a factor of 10. These scaled scores (representing the two dimensions) served as the output, while the corresponding text and the expert-defined criteria used as reference formed the input.

During the development phase, we adopted a small-sample fine-tuning strategy to customize GPT-4o for our domain-specific application. This approach, which leverages a relatively limited number of high-quality training examples, has been shown to be both efficient and practically effective in enhancing model performance on specialized tasks [5]. We prepared and uploaded the curated dataset through OpenAI's official platform and used their fine-tuning API to tailor GPT-4o. The resulting customized model served as the backbone of our scoring system.

[5]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

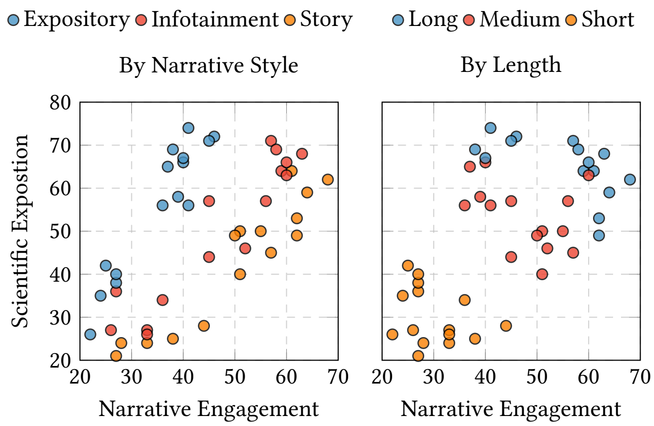

Fig. 10. Each point represents one of 45 science communication texts, plotted by its average audience rating for narrative engagement (x-axis) and scientific exposition (y-axis), based on 27 crowd-sourced rubric-based evaluations per text. The left panel groups texts by narrative style: Expository (informational, fact-focused), Story (highly narrative), and infotainment (represents infotainment-style revisions that blend factual exposition with narrative strategies). The right panel groups texts by length (Short=50 words, Medium=150 words, Long=300 words).

Technical Evaluation To validate the reliability of this scoring mechanism, we conducted a formal evaluation. We constructed a controlled dataset consisting of five source articles, each systematically rewritten into three different lengths (long, medium, short) and expressed in three different styles (expository, medium, story). This design yields nine distinct variants per article, resulting in a total of 45 text samples. From this dataset, we randomly selected 33 samples for fine-tuning GPT-4o, while reserving 12 samples for evaluation. The fine-tuned model was assessed against human ratings on two key dimensions: narrative engagement and scientific exposition. On the held-out test set, the fine-tuned model demonstrated a high degree of alignment with human judgment, achieving Pearson correlation coefficients of 0.90 and 0.91 for narrative and exposition scores, respectively. In addition, the model's predictive reliability was reflected in RMSE values of 6.48 and 7.02. These results indicate that the fine-tuned LLM scoring mechanism can effectively approximate human evaluative patterns, thereby providing a reliable and scalable alternative to manual scoring.

A.2.1 Example of Content.

Please view the materials via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/7V7gS5xcQdZwo0mLeBbfiQe6HEgU+02HqdaupBV9tA0/

A.2.2 Survey used for gathering audience feedback.

Please view the survey via the anonymous link: https://cryptpad.fr/doc/#/2/doc/view/XfWs-wD3qmBXSnEC0YqM9EZg2GO++H2RJYUqyrcvj1I/

A.2.3 Score Criteria.

Please view the criteria via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/uNMusLpCPWGwzqKWi04F0TY+20nW2hnG1NkS1V2BHB4/

A.3 Materials used for experiment

Please view the materials via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/Q3Jhj+HhzHtt9zYqyF0Sv4mziQYBp6oWl43a84Gqme

A.4 Survey

Part 1: Metacognition

Metacognitive Knowledge: This pertains to an individual's awareness and understanding of their own cognitive processes and strategies

Q1: I am aware of my writing goals during the editing process.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Metacognitive Regulation: This involves the active management of one's cognitive processes through planning, monitoring, and evaluating

Q2: I set specific goals for what I wanted the narrative to achieve.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q3: I reflect on my writing strategies or editing choices while using the AI writing tool. (Indicates real-time assessment of strategy effectiveness.)

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q4: During writing, I regularly checked whether the narrative was staying on track with my intended message.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q5: I can clearly identify areas of my writing that need improvement when using the AI tool.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q6: After writing, I reviewed the narrative to assess how well it communicated the scientific content.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q7: I am able to adjust my writing strategies during the editing process.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 2: Control (Control:)

Q8: I felt in control of the writing process while interacting with the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q9: I was able to override or ignore the system's suggestions when I thought it was necessary.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q10: I determined the direction and flow of the science narrative, not the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 3: Autonomy (Autonomy:)

Q11: I felt free to make my own choices during the co-writing process with the system.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q12: The system supported my ability to express my own ideas in the narrative.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q13: I did not feel pressured to accept the system's suggestions.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

A.5 Participants demographic information

ID	Age	Gender	Education	AI Writing Use	Writing Confidence	Occupation
1	26	Male	Postgraduate	Occasionally	Confident	(a)
2	27	Male	Postgraduate	Daily	Confident	(a), (b), (c), (d)
3	26	Male	Postgraduate	Daily	Confident	(b), (d)
4	25	Female	Postgraduate	Daily	Confident	(a), (b), (c)
5	24	Male	Postgraduate	Daily	Confident	(a)
6	28	Female	Postgraduate	Weekly	Neutral	(a)
8	28	Male	Postgraduate	Occasionally	Neutral	(a)
7	29	Female	Higher than postgraduate	Daily	Confident	(a), (b)
9	31	Male	Postgraduate	Weekly	Neutral	(a)
10	24	Female	Postgraduate	Occasionally	Confident	(a), (c)
11	29	Female	Postgraduate	Weekly	Neutral	(a)
12	26	Male	Postgraduate	Weekly	Neutral	(a)
14	27	Male	Postgraduate	Daily	confident	(a), (b)
15	24	Female	Postgraduate	Weekly	Neutral	(a)
16	30	Male	Postgraduate	Weekly	Neutral	(a)

Occupation: (a) PhD Student / Postdoctoral Researcher / University Faculty / Researcher;
(b) Science Journalist / Media Producer;
(c) Educator / Teacher;
(d) Online Science Content Creator (e.g., YouTube, Blog, TikTok, etc.)

A.6 User Study Results

1. Visualization of interaction behaviors from 16 participants across two revision directions:

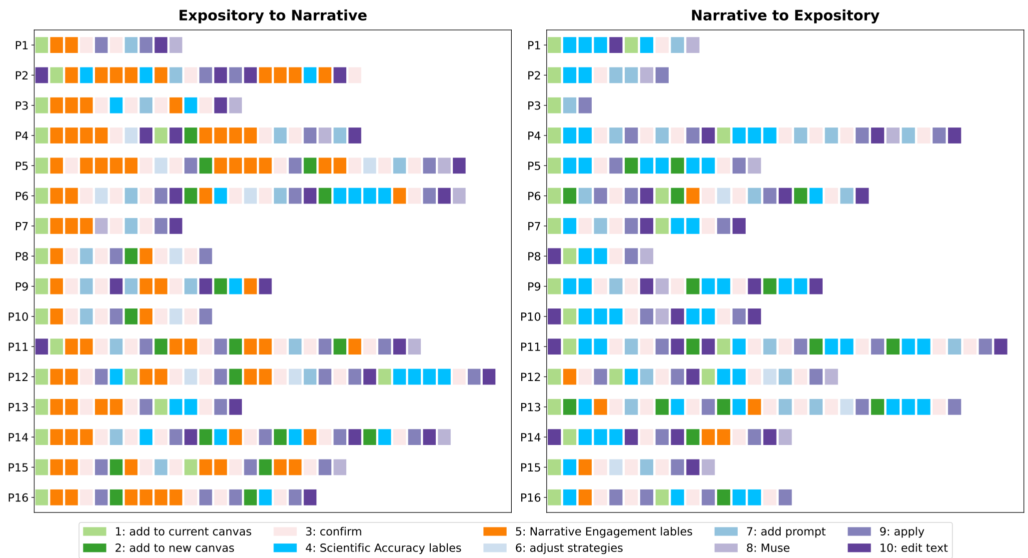

Fig. 11. Visualization of interaction behaviors from 16 participants across two revision directions.

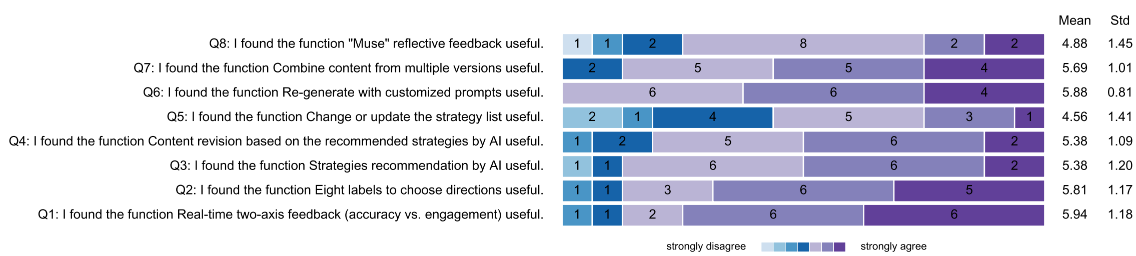

Fig. 12. Functional Evaluation of SpatialBalancing.

A.7 Prompts

A.7.1 *Recommender.*

The blue word will be replaced by input information.

```
# Base prompt
You are an expert in science communication narrative text revision and strategy recommendation.
Your task is to analyze the given text and recommend effective strategies to improve it.

# Order prompt
Step 1: Analyze the Text.
Position: Identify where the selected text {text} appears in the {overall_content}.
Granularity: Determine whether the text consists of sentences, paragraphs, or a complete document.
Core Message: Extract the key ideas that must be preserved and effectively conveyed in text.

Step 2: Select Strategies Review the available strategy list {strategy_info}, including their
definitions, examples, and usage instructions. Choose a set of strategies that align with the
text's characteristics and modification goals. Ensure the selected strategies are compatible
when combined. Consider multiple ways to apply the strategies for improvement.
Only choose strategies mentioned above, and use them appropriately.
Provide {generated_number} different versions, each using distinct or complementary strategy sets.
These different versions should use different strategies, preferably with varied combinations of
strategies.

Step 3: Output the Strategy List Return the strategy selection in JSON format with multiple versions:
{
"Version1": [ "Strategy_A", "Strategy_H", "Strategy_J", "Strategy_B"],
"Version2": [ "Strategy_F",..., "Strategy_E"],
...,
"Version_number": [ "Strategy_G", "Strategy_M",..., "Strategy_C",...,"Strategy_D"]
}
Do not include any extra commentary or explanation outside the JSON.
Let's think step by step.
```

A.7.2 *Generator.*

The blue word will be replaced by input information.

Generate new text based on user selected goals

```
# Order prompt
You are an expert in science communication narrative strategy. Your task is to revise the
given text using the recommended strategies and provide a concise overview of how the
strategies were applied.

Step 1: Review the Strategy List
- Read the strategy list {strategy_info}, including each strategy's definition and
how it is typically used.

Step 2: Apply all the Strategies mentioned in the strategy list to the Text: {text}.
Even if the original text already contains elements that align with the strategy, enhance it further
based on how the strategy should be applied.
Also, consider the position of the given text in the whole context {overall_content}.
Make the changed text coherent with the context.

Step 3: Summarize the Application
- Summarize how each selected strategy was applied.
- Keep the summary concise and short to indicate what specific changes have been made using
separate strategies.

Step 4: Do not omit or alter any important information from the original text, but ensure that the
generated text is distinct from the original.

Step 5: If the content is primarily narrative in nature, supplement it with scientifically grounded
explanations, relevant data, or reliable sources to enhance credibility and depth.

Step 6: Output the Result Return a JSON with the following structure:
{
"strategies": ["Strategy_A", ..., "Strategy_B", "Strategy_C", "Strategy_D"],
"summary": "Summarize how each strategy was applied and what specific changes were made to the content
             based on each strategy. Example: Changed 'Photosynthesis is the process plants use to
             make food.' to 'What if plants could teach us how to turn sunlight into fuel?
             Focus only on the changes from the previous version.'",
"newText": "Modified version of the text. Even if the original text already contains elements that
             align with the strategy, enhance it further based on how the strategy should be applied."
}

Do not include any extra commentary or explanation outside the JSON.
Let's think step and step.
```

A.7.3 Scorer.

The blue word will be replaced by input information.

```
# Base prompt
You are an engaging audience for science communication.
Given a narrative, evaluate it on two dimensions: (1) Narrative Engagement and (2) Scientific Exposition.
using the detailed scoring rubrics below.
Provide a numerical score from 0 to 100 for each dimension, along with a brief explanation justifying
your rating.

Dimension 1:
Narrative Engagement: Evaluate how effectively the narrative captures attention, evokes emotion,
sparks curiosity, and maintains reader engagement.
Scoring Rubric:
0-20: Extremely boring and dry, no storytelling elements,
21-40: Barely engaging, logical but lacks emotion or creativity,
41-60: Moderately engaging, uses some analogies or description but still feels academic,
61-80: Quite engaging, includes storytelling techniques and relatable examples,
81-100: Highly immersive, vivid storytelling with strong emotional or narrative appeal.


Dimension 2: Scientific Exposition: Assess how well the narrative explains scientific concepts with
clarity,
correctness, and alignment with established knowledge.
Scoring Rubric:
0-20: Highly inaccurate or pseudoscientific, major factual errors,
21-40: Misleading or speculative, lacks clarity or evidence,
41-60: Mostly accurate but vague or oversimplified,
61-80: Generally accurate, minor imprecision, lacks citations,
81-100: Highly accurate, precise, and well-aligned with scientific consensus.

# Order prompt
This is the original text: {text} and its score {currentScore}. Please use this as a reference.
Compare the current version with the original one in terms of scientific exposition and narrative
engagement, and assess whether it performs better or worse than the previous version.
Compared to the previous version's scores, assign a score difference within a reasonable range.
```