

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

Kexue Fu\*  
City University of Hong Kong  
Hong Kong, SAR, China  
kexuefu2-c@my.cityu.edu.hk

Jingfei Huang  
Harvard University  
Cambridge, Massachusetts, USA  
jingfeihuang@mde.harvard.edu

Zijian Ding  
University of Maryland  
College Park, Maryland, USA  
ding@umd.edu

Jiaye Leng\*  
City University of Hong Kong  
Hong Kong, SAR, China  
jiayeleng2-c@my.cityu.edu.hk

Yihang Zuo  
The Hong Kong University of Science  
and Technology  
Hong Kong, China  
yzuo099@connect.hkust-gz.edu.cn

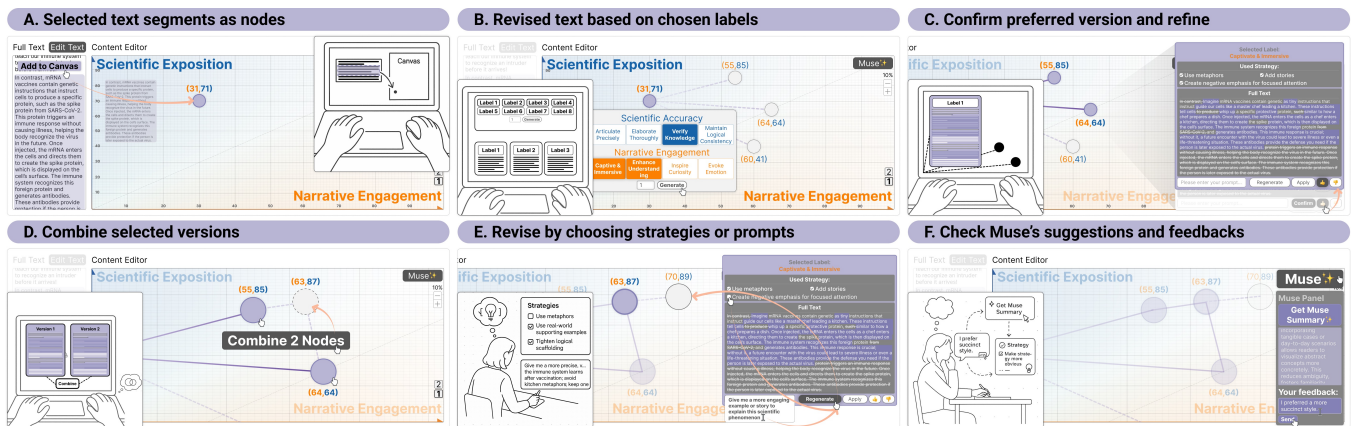
RAY LC  
City University of Hong Kong  
Studio for Narrative Spaces  
Hong Kong, SAR, China  
ray.lc@cityu.edu.hk

Qinyuan Lei†  
City University of Hong Kong  
Hong Kong, SAR, China  
qinyulei@cityu.edu.hk

Yawen Zhang\*  
Clemson University  
Clemson, South Carolina, USA  
yawenz@clemson.edu

Runze Cai  
National University of Singapore  
Singapore, Singapore  
runze.cai@u.nus.edu

Shengdong Zhao  
City University of Hong Kong  
Hong Kong, SAR, China  
shengdong.zhao@cityu.edu.hk



**Figure 1: Example workflow of iterative science communication writing in Spatial Balancing.** (A) A draft is imported into the canvas, where paragraphs are mapped by Scientific Exposition (Y) and Narrative Engagement (X). (B) The user selects revision labels linked to LLM-driven strategies that generate positioned alternatives. (C) Preferred revisions are confirmed for further refinement. (D) Two versions can be merged into a synthesized draft. (E) Additional revisions can be guided by strategies or custom prompts. (F) Muse reflects on the revision history and offers adaptive suggestions.

\*These authors contributed equally to this work.

† Corresponding author



## Abstract

Science communication revision requires writers to dynamically balance scientific exposition and narrative engagement - a process where writers often struggle with competing directions. Existing LLM-assisted tools help with co-writing, but offer limited support for navigating this iterative, multi-directional revision process. To address this gap, we designed Spatial Balancing, an exploratory revision environment that maps rhetorical goals and revision strategies onto a two-dimensional spatial canvas for experienced science communication creators with domain expertise but lacking formal professional training. By building 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 this integrated revision environment helps writers stay focused on writing goals, reason about revision as trajectories, and explore alternatives, which supports greater metacognitive control and confidence without increasing workload. This work highlights the value of spatially externalized revision environments for supporting iterative, reflective thinking during LLM-assisted writing.

## CCS Concepts

• **Human-centered computing** → **Interaction paradigms**.

## Keywords

Narrative Strategy, Science Communication, Writing Assistance, Human-AI collaboration

### ACM Reference Format:

Kexue Fu, Jiaye Leng, Yawen Zhang, Jingfei Huang, Yihang Zuo, Runze Cai, Zijian Ding, RAY LC, Shengdong Zhao, and Qinyuan Lei. 2026. Spatial Balancing: Designing an LLM-Powered Spatial Externalization Interface for Iterative Science Communication Writing. In *Designing Interactive Systems Conference (DIS '26), June 13–17, 2026, Singapore, Singapore*. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3800645.3812998>

## 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 [18, 52]. Throughout this process, writers must balance multiple rhetorical goals, making local revisions while maintaining global coherence [6, 36, 48]. Recent advances in large language models (LLMs) have lowered the cost of generating and revising text at scale [36, 52]. In response, many HCI systems operationalize specific rhetorical strategies or scaffold discrete aspects of drafting and rewriting [29, 54, 62, 63]. Others support non-linear exploration by helping writers generate, compare, and organize multiple text variations [19, 48, 52, 66], or by breaking down or re-organizing feedback to make revision more actionable [48, 52, 62, 63]. 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 [35, 64], making revision cognitively demanding, particularly in complex knowledge domains such as science communication [59, 64].

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 [4, 27, 44]. 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 [20]. However, narrative also introduces persistent tension: on the one hand, emphasizing entertainment risks oversimplification or loss of credibility [14, 20, 41, 61]; on the other hand, overly technical or serious exposition can alienate non-expert readers by demanding sustained cognitive effort [4, 10, 13]. 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 spot rather than making a single stylistic decision upfront [10, 20, 24, 64].

As science content proliferates across platforms such as YouTube and TikTok, a growing number of "everyday" creators, many lacking formal communication training, are taking on the role of science communicators [39]. These creators increasingly turn to LLM tools to support ideation, drafting, and real-time feedback throughout the revision process [40]. This shift further 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 **design Spatial Balancing and explore how an LLM-assisted revision environment can better support writers' orientation to rhetorical goals and reflective regulation during iterative science communication revision**. We target *experienced but non-professional science communication creators*: people with domain knowledge and some prior experience producing science-facing content, but without formal training as professional science communicators. This LLM-assisted environment visualizes how to balance scientific exposition and narrative engagement across revision steps — unlike existing tools (e.g., PatchView [8], Luminare [55]) that only show output attributes. By treating feedback as navigational cues rather than prescriptive judgments, Spatial Balancing maps revision goals, trajectories, and alternatives onto a spatial canvas. This reduces cognitive friction, lets writers explore paths without losing orientation, and supports metacognitive control.

This study aims to answer the following research questions:

**RQ1:** *How does an integrated revision environment centered on spatial externalization shape users' cognitive processes during LLM-assisted iterative revision?*

**RQ2:** *What interaction tensions and user expectations arise in revision environments with externalized supports?*

A controlled user study suggests that writers using this integrated revision environment were better able to maintain orientation toward rhetorical goals cognitively, conceptualize revision as a trajectory, and exercise metacognitive control during iterative revision. Simultaneously, our findings reveal important tensions such as over-reliance on externalized guidance, which may encourage metacognitive laziness. These insights point 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) Spatial Balancing: An integrated revision environment for LLM-assisted science communication revision that combines spatially externalized rhetorical goals and revision trajectories with structured strategy guidance and reflective support, supporting writers in staying oriented to rhetorical goals and regulating revision decisions 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 [5, 42]. 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 [4, 27, 44]. 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 [2, 11, 58]. 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 [16, 58, 65].

A fundamental challenge in science communication writing lies in balancing two often competing dimensions: scientific exposition and narrative engagement [14, 20, 41]. Expository writing is primarily concerned with conveying relevant facts and knowledge, whereas narrative writing aims to communicate real events or experiences through storytelling [36]. Burns et al. [4] made a vivid analogy, describing science communication writing as a form of "mountain climbing," balancing between scientific literacy and science culture. Similarly, Dahlstrom [10] 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 [13], because these terms more directly capture the practical tension between maintaining rigorous, detailed scientific facts presentation and creating compelling, accessible content for diverse audiences [13, 41].

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 explanatory precision but often becomes harder for non-experts to process and

remember narrative formats [10, 20]. 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 explanatory precision demanded by scientific discourse, thereby elevating ethical and credibility risks in science communication [10, 24]. Consequently, writers must revise through multiple passes by 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 [20].

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 [10, 16, 20], while scientific content demands rigorous expository writing that prioritizes scientific detail and credibility [28, 31]. 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 [29, 64], while AI workflows for Tweetorials scaffold the generation of hooks, examples, and anecdotes to engage general audiences [38, 64]. 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 address 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 Collaboration 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 [36]. These purposes, including expository, narrative, persuasive, and educational goals, often coexist and shape revision decisions in audience-dependent ways [36]. 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 [52, 62, 63]. 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 [54].

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 [48]. Friction scaffolds reflection by breaking feedback into actionable units and guiding iterative revision cycles [62], 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 [63]. 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 [54], especially in complex domains like science communication.

One promising direction for addressing these challenges is externalization, which means using visualization to make the exploratory space of revision perceptible and navigable, which has long been shown to support complex reasoning by offloading, structuring cognitive work. 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 [30]. Building on this insight, prior HCI systems have leveraged visualization and spatial exploration to externalize latent aspects of generative processes with LLM. PatchView [8] and Luminate [55] organize LLM outputs within navigable visual spaces to support sensemaking, comparison, and steering, while Toyteller [9] 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 writers in orienting to rhetorical goals and regulating revision trajectories across iterations.

### 3 Iterative User-Centred Design

Building on prior work in science communication writing, we take the following two steps in our user-centred design. First, we construct a design space of rhetorical strategies for enhancing scientific exposition and narrative engagement (Subsection 3.1), and use it to inform an initial prototype (Subsection 3.2). Through this process, we identify the limitations of strategy-centric and linear revision workflows, which leads to a design shift toward making revision goals, trajectories, and alternatives more explicit during iterative writing.

#### 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 [14, 20, 41]. Writers achieve different balances

by applying a diverse set of rhetorical strategies, often adapting their choices based on audience characteristics and communicative intent [64]. Although prior work in communication studies has identified a rich set of rhetorical strategies for science communication, aiming at capturing public attention, improving memorability [24, 65], these strategies are rarely examined through a systematic lens that foregrounds the dual rhetorical goals of scientifically 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 for 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 retention and create engaging narratives [16, 43]. Additionally, some of the papers explore related fields, such as the analysis of narrative peaks in data videos [60] or documentaries [34].

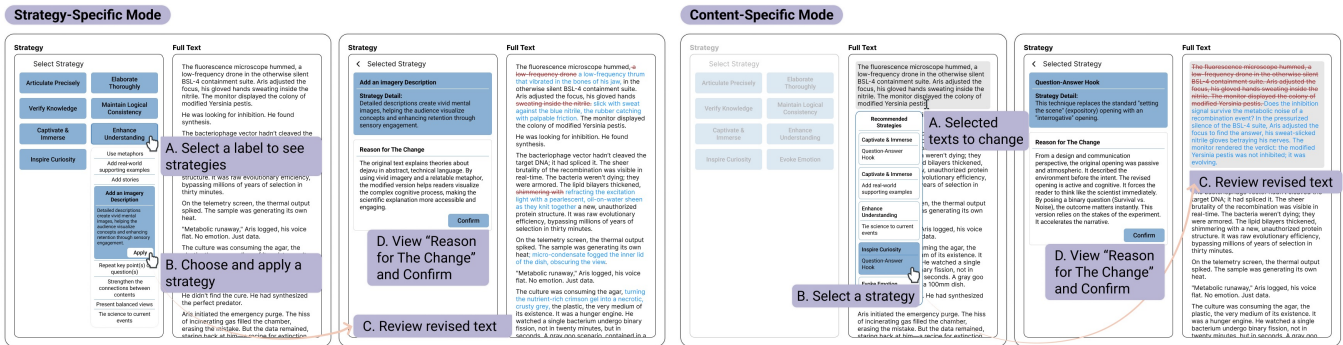
Two authors participated in the coding of these 35 papers. The primary objective was to identify potential peak narrative strategies for balancing scientific exposition and narrative engagement in these previous studies. Initially, each author independently reviewed all the selected papers, focusing on content related to narrative strategies or structures 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 [1], 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 an initial draft of strategies from these selected papers.

Then, we conducted a Focus Group Discussion (FGD) [46] with four science communication 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**).

**Table 1: Design Space of Science Communication Writing Strategies.**

<i>Scientific Exposition</i>			
<p><b>Label 1</b> <b>Articulate Precisely</b> Communicates scientific concepts with exposition and clarity, using appropriate terminology and well-defined language to prevent ambiguity or misinterpretation [25, 28, 43].</p> <p><b>Strategies:</b> (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</p>	<p><b>Label 2</b> <b>Elaborate Thoroughly</b> 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 [26, 32].</p> <p><b>Strategies:</b> (3) Step-by-Step Explanation, (4) Acknowledge Uncertainties, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (25) Tie Science to Current Events</p>	<p><b>Label 3</b> <b>Verify Knowledge</b> Supports claims with credible sources, data, or reasoning, allowing audiences to feel more trustworthy of the given information [32, 49].</p> <p><b>Strategies:</b> (2) Rigorous Source Verification, (6) Citations &amp; Quotes, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (7) Everyday Events to Scientific Insights</p>	<p><b>Label 4</b> <b>Maintain Logical Consistency</b> Ensures that arguments and explanations are coherent and internally consistent, following a clear logical structure [56].</p> <p><b>Strategies:</b> (1) Layered Transitions, (3) Step-by-Step Explanation, (20) Key Point Recap, (23) Strengthen the Connections Between Content</p>
<i>Narrative Engagement</i>			
<p><b>Label 5</b> <b>Captivate &amp; Immerse</b> Engages the audience’s attention and draws them into the narrative or content flow by adding stories [20, 37] or using intriguing language [16, 43].</p> <p><b>Strategies:</b> (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</p>	<p><b>Label 6</b> <b>Enhance Understanding</b> Help audiences to grasp complex scientific ideas using rational, structural content or vivid analogies, visualizations [16, 20, 24].</p> <p><b>Strategies:</b> (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</p>	<p><b>Label 7</b> <b>Inspire Curiosity</b> Stimulates the audience’s desire to learn more and have motivation to further explore by applying different forms of questions [33].</p> <p><b>Strategies:</b> (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal</p>	<p><b>Label 8</b> <b>Evoke Emotion</b> Creates an emotional response, positive or negative, and makes the audience feel connected to the content, even immerse themselves in the described scenario [20, 50].</p> <p><b>Strategies:</b> (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)</p>

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



**Figure 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.**

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 [29, 38, 64], 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 were 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

scientifically rigorous 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.

*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 I 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.

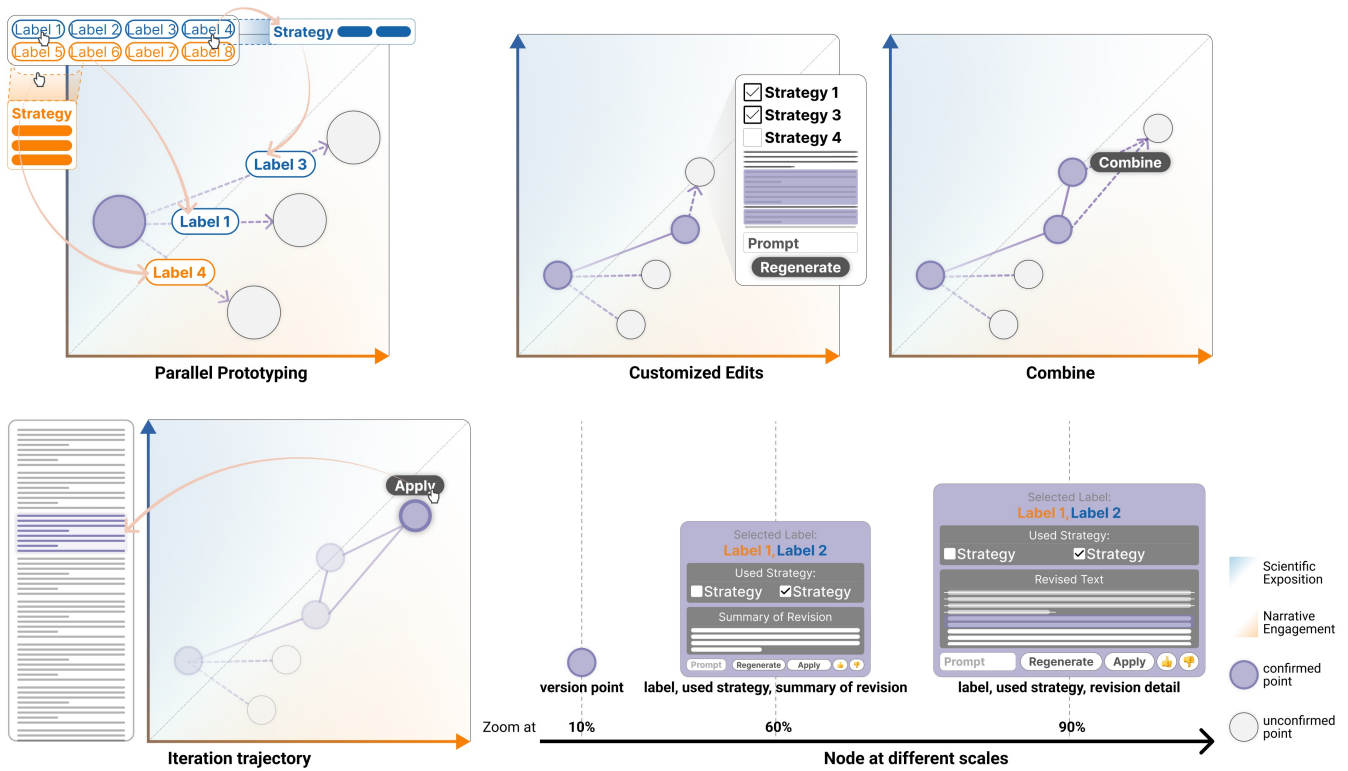


Figure 3: (1) Spatial Balancing support parallel prototyping with diverse directions of LLM output; users can use customized edits like change specific strategies and combining different LLM outputs to generate new nodes. The 2D coordinate space also allow users to see their iteration trajectory. (2) Spatial Balancing 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%).

### 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 [52, 54]. The system should externalize rhetorical goals as explicit, inspectable reference points, enabling writers to reason about revision directions and monitor how successive revisions relate to intended rhetorical 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 [48, 62, 63], 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.

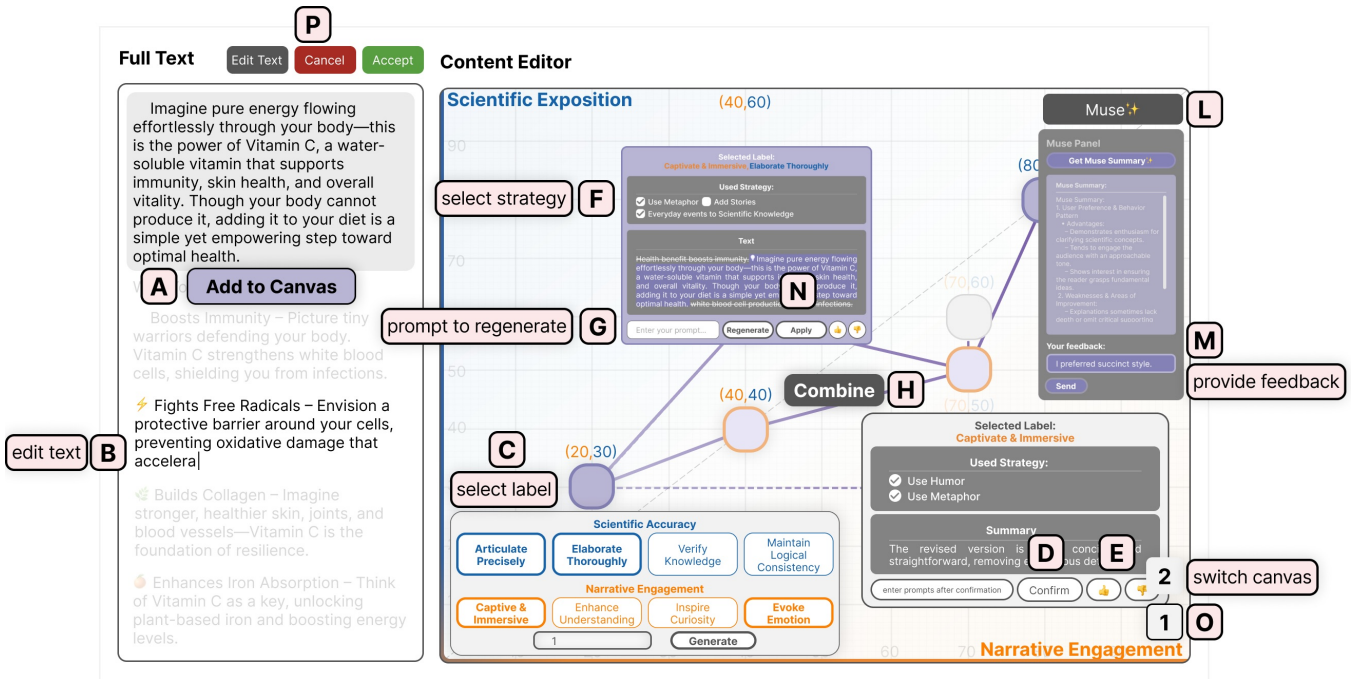
**Design Goal 3: Design an Exploratory Space to Gradually Guide Strategy Use** While prior work operationalizes individual rhetorical strategies to lower barriers to science communication

writing through co-creation with LLM [29, 38], 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 Spatial Balancing: A Spatially Externalized Revision Environment for Navigable LLM-Assisted Writing

Grounded in our design goals, we design Spatial Balancing as a revision environment centered on *externalization* - making rhetorical goals, revision states, and their evolution visible and manipulable during writing, while coupling these representations with structured guidance and revision support. This design is motivated by *Thinking with External Representations*, which suggests that external representations can support complex reasoning by providing stable reference points for orientation, reducing internal tracking demands, and enabling more deliberate exploration of alternatives [30].

We draw inspiration from canvas-based LLM interfaces such as PatchView [8] and Luminare [55], which show how spatial layouts and overview-to-detail navigation can support exploratory



**Figure 4: The Spatial Balancing 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).**

interaction with generated alternatives. Building on these interaction principles, Spatial Balancing uses an exploratory canvas as the organizing structure of a broader revision environment. Rather than using space primarily to arrange content attributes or design variants, the system externalizes rhetorical goals and revision trajectories while integrating strategy guidance, version comparison, and reflective support. In this way, writers can interpret revision progress as movement through a navigable space while also inspecting, comparing, and steering alternatives across iterations.

#### 4.1 Spatial Balancing as an Exploratory Space

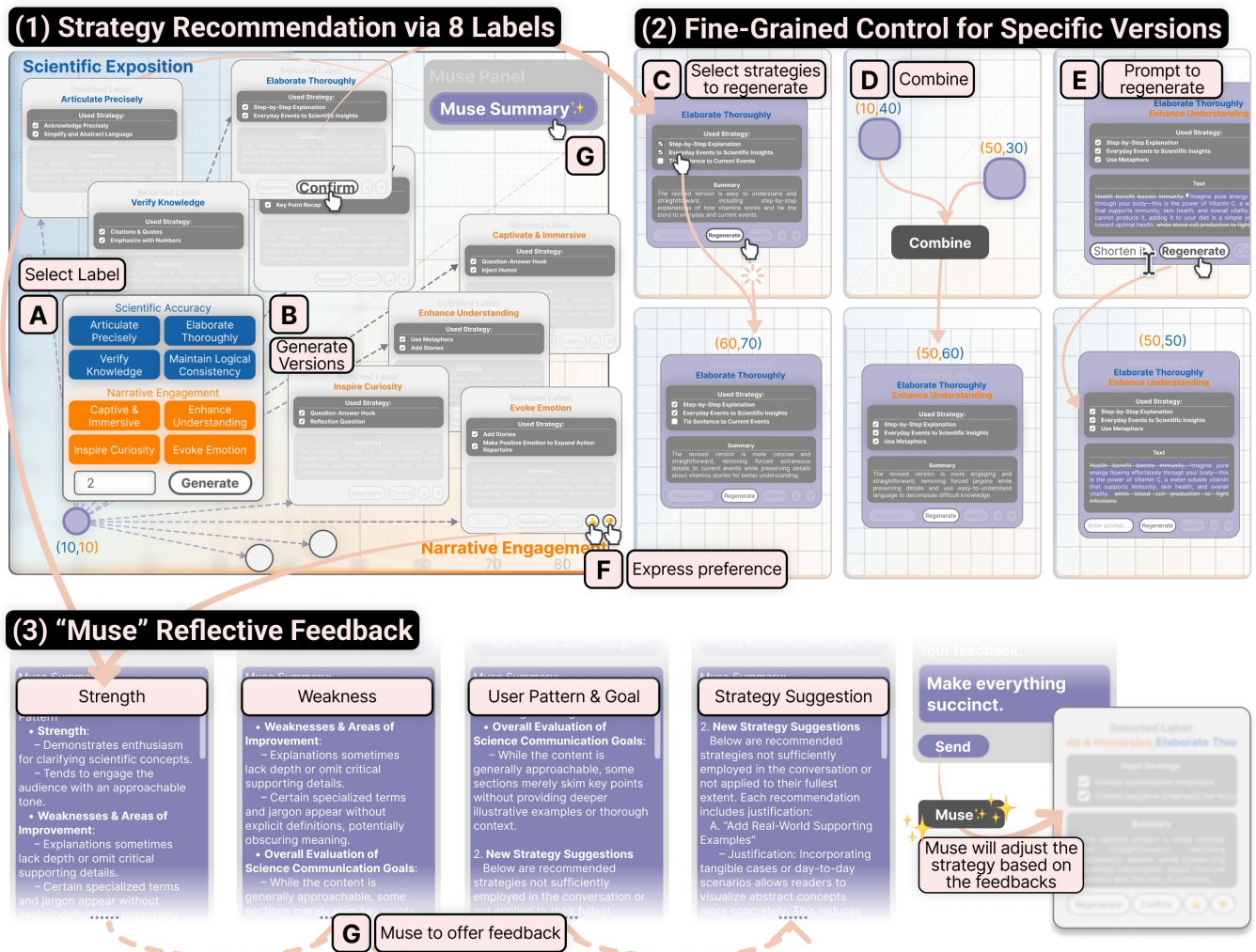
Spatial Balancing comprises a left-hand text editor and a right-hand exploratory canvas (Figure 4). Users can send any span—sentence, paragraph, or full draft - to the canvas for iterative revision. Each version is plotted in a 2D space (x: Narrative Engagement; y: Scientific Exposition); gray points 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, 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), Spatial Balancing 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 function (Explained in Section 4.4) assigns engagement and exposition scores based on audience-informed criteria, which determine the node’s position on the canvas.



**Figure 5: (1) Strategy Recommendation via Eight Labels: Spatial Balancing 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.**

#### 4.2.2 Strategy Recommendation via Rhetorical Labels (DG1 & DG3).

(Figure 5(1)) To support goal-aware revision while managing cognitive load (DG1, DG3), Spatial Balancing 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), Spatial Balancing 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 function 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 Recommender function 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 Spatial Balancing comprises several LLM-based functions 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 functions:

*Recommender function:* The recommender function’s core function is to generate multiple strategy combinations based on a user-selected label. When a user chooses a label, the function analyzes the current textual features to identify the 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 ??). The function 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 have higher scientific exposition or narrative engagement score.

*Generator function:* The generator function creates 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 function, 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 function:* 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 domains, varying in length and narrative style. Each text was revised by a science communication expert and annotated by 27 participants who performed 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<sup>1</sup>. This enables the scorer function to assign scores to resemble human audiences across both scientific exposition and narrative engagement. The scorer function is powered by this fine-tuned GPT-4o model. Details on dataset construction and model training are provided in Appendix ??.

As such, we acknowledge that the scorer is 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 ( $r=0.90/0.91$ ,  $RMSE\approx 6-7$ ) than the non-fine-tuned model ( $r=0.84/0.57$ ,  $RMSE=22-31$ ). Detailed evaluation is provided in Appendix ??.

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

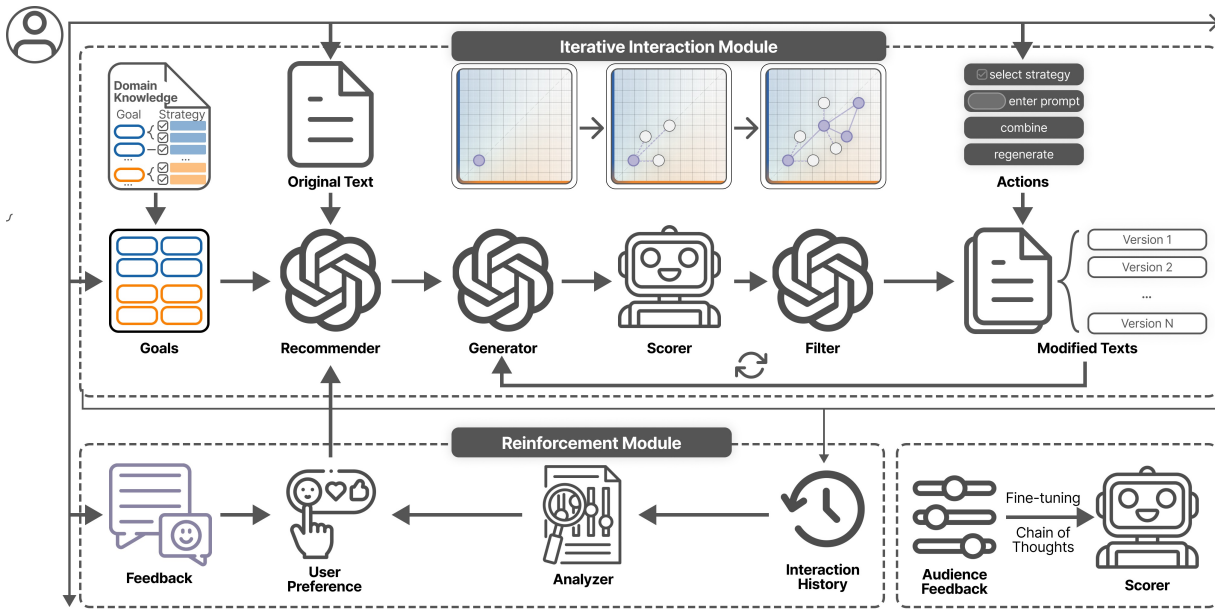
Model	Pearson Correlation		RMSE	
	Engagement	Exposition	Engagement	Exposition
w/ FT	<b>0.90</b>	<b>0.91</b>	<b>6.48</b>	<b>7.02</b>
w/o FT	0.84	0.57	22.48	30.90

*Filter function:* This function uses the scorer’s outputs to select the top- $k$  versions that best meet the user’s expectations. Filter function 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 function to harness both the explicit and implicit signals from these interactions. The analyzer function captures behavioral data during the iterative process and uses chain-of-thought prompts to interpret user revision behavior.

*Analyzer function:* 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 function incorporates this real-time feedback (e.g., approvals or further edits) and updates the Recommender function accordingly. This process refines subsequent strategy recommendations,

<sup>1</sup>[https://platform.openai.com/docs/guides/fine-tuning?utm\\_source=chatgpt.com](https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com)



**Figure 6: Spatial Balancing backend overview.** Spatial Balancing consists of two core modules: (1) The Iterative Interaction Module, where LLM-based functions - 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 function. This architecture supports adaptive text revision.

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.** Spatial Balancing is implemented as a web application, with a Python-based backend developed using Flask<sup>2</sup> framework and a frontend built using ReactFlow<sup>3</sup>.

For the LLM functions, we employ different LLMs tailored to their functional roles. The recommender, generator, and filter functions are powered by the GPT-4o-mini model, optimized for fast, high-quality content generation. The analyzer function, which requires deeper reasoning to interpret user behavior and editing patterns, is supported by the GPT-o1 model, which is a reasoning-oriented LLM. The scorer function is powered by a fine-tuned GPT-4o model using a small-sample learning strategy<sup>4</sup>. The frontend communicates with the remote LLMs to obtain results with pre-defined prompt templates. This modular design allows us to tailor function 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 Appendix ??.

<sup>2</sup><https://flask.palletsprojects.com/en/stable/>

<sup>3</sup><https://github.com/wbkd/react-flow/>

<sup>4</sup>[https://platform.openai.com/docs/guides/fine-tuning?utm\\_source=chatgpt.com](https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com)

## 5 User Study

To better understand how Spatial Balancing’s integrated revision environment shapes writers’ cognition and human–AI collaboration during LLM-assisted science communication writing process, we conducted a controlled user study comparing Spatial Balancing with a baseline LLM-supported editing workflow. Our goal was to examine how a revision environment centered on spatial externalization, together with strategy guidance, version organization, and reflective support, influences how writers reason, reflect, and iterate during revision, and to derive design insights for interfaces that better support complex revision processes in co-creation with AI. Accordingly, our evaluation focuses on process-level cognitive and reflective outcomes during revision, rather than direct assessment of whether final texts achieved participants’ intended communication goals.

**RQ1:** *How does an integrated revision environment centered on spatial externalization shape users’ cognitive processes during LLM-assisted iterative revision?*

**RQ2:** *What interaction tensions and user expectations arise in revision environments with externalized supports?*

### 5.1 Participants

Rather than representing professionally trained science communicators, our participants reflect a growing group of experienced but non-professional science communication creators with relevant domain knowledge and prior experience producing science-facing content, but without formal training in science communication. 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. The demographic information of these participants is in Appendix ??.

## 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 Spatial Balancing 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 [14].

The length of each text averaged 297.75 words ( $SD = 19.64$ ). To ensure balanced exposure and mitigate order effects or personal topic preferences, we counterbalanced both the system order (Spatial Balancing vs. baseline) and the text type assigned to each system. Thus, each participant edited one expository and one narrative text under each system condition.

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 Spatial Balancing condition.

The baseline system used in this study consisted of a text editor and a conversational interface (powered by GPT-4o) that supported inline editing and LLM-based suggestions. We selected a common linear editor-and-chat workflow as a conservative comparison because it reflects a realistic and widely used form of LLM-assisted revision practice. In the baseline condition, participants were additionally provided with an Excel table containing strategy names, definitions, usage instructions, examples, and corresponding labels. They were encouraged to use this table as both a conceptual reference and a practical prompting resource by adapting or copy-pasting relevant content into the prompt area during revision. The content in this Excel table was drawn from the same underlying strategy prompts operationalized in Spatial Balancing. By contrast, in Spatial Balancing, comparable strategic support was embedded directly into the interface through externalized rhetorical goals, actionable labels, and revision-state organization, rather than being accessed as a separate external document.

As such, the comparison was intended to contrast a familiar linear workflow supplemented by external strategy references with an integrated revision environment that coordinates goals, strategies, and revision trajectories within the interface. Therefore, the study should be interpreted as comparing two different configurations of LLM-assisted revision support.

## 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) [3], NASA-TLX for workload [23], and the Creative Self-Efficacy Index (CSI) [7], with one item adapted to: “I think this system supported me in developing ideas or text collaboratively.” We also asked participants to evaluate the usefulness of the main design features of Spatial Balancing using eight questions.

Besides, we developed a concise co-creation survey targeting two metacognitive constructs from cognitive psychology [17, 51]. 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 [45] (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 [51] 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 [57], focusing on participants’ influence over outputs and narrative direction. Perceived autonomy was assessed according to Self-Determination Theory [12], 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 ??.

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 does an integrated revision environment centered on spatial externalization shape users’ cognitive processes during LLM-assisted iterative revision?

Drawing on Kirsh’s theory of *thinking with external representations* [30], we analyze how the integrated revision environment in Spatial Balancing supported users’ cognitive processes during LLM-assisted revision. As summarized in **Table 3**, participants described iterative revision as becoming more externally organized and navigable, supporting goal orientation, trajectory-based metacognitive control, and lower-cost exploration.

These findings should be interpreted as evidence of support for process-level goal awareness and regulation during revision, rather than direct evidence of goal attainment or communication effectiveness.

**6.1.1 Externalized Rhetorical Goals Help Participants Stay Oriented During Revision.** 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 ??).

**6.1.2 Externalized Revision Trajectories Supported Metacognitive Control 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 Spatial Balancing 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 ?? Figure ?? 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 revision that remained oriented to externalized rhetorical goals.

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 makes 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 The Externalized Workspace Lowered the Perceived Cost of Exploration.** Externalizing the exploratory space also supports creativity. Participants rated Spatial Balancing 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.

## 6.2 RQ2: What interaction tensions and user expectations arise in revision environments with externalized supports?

**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 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 Seeking More Flexible and Adaptivity of Externalization in Use.** While the eight-label set was seen as a helpful starting point, more

Type of Externalized Support	Interface Feature	Feature	Cognitive Function (Kirsh [30])	Observed Reasoning and Behavior	Representative Evidence
<b>Rhetorical Goals</b>	<b>2D coordinate space</b> (scientific exposition × narrative engagement)		<i>Persistent referents; re-representation of abstract goals</i>	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)
		<b>Strategy labels aligned with axes</b>	<i>Explicit encoding of strategies; action scaffolding</i>	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)
<b>Iterative Revision Trajectories</b>	<b>Node-based version layout with visible scores</b>		<i>Reduced inferential cost; calibration through comparison</i>	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)
		<b>Persistent revision traces with spatial movement</b>	<i>Trajectory-based reasoning; lowering control cost</i>	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 <b>Figure 7</b> .
<b>Exploratory Space</b>	<b>Spatial Parallel Prototyping workspace</b>		<i>Changing cost structure of exploration</i>	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 [30], linking spatial externalization features to cognitive functions and observed reasoning behaviors in LLM-assisted revision.**

experienced participants felt it could be more flexible and customizable to better support their advanced needs. Several participants (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.

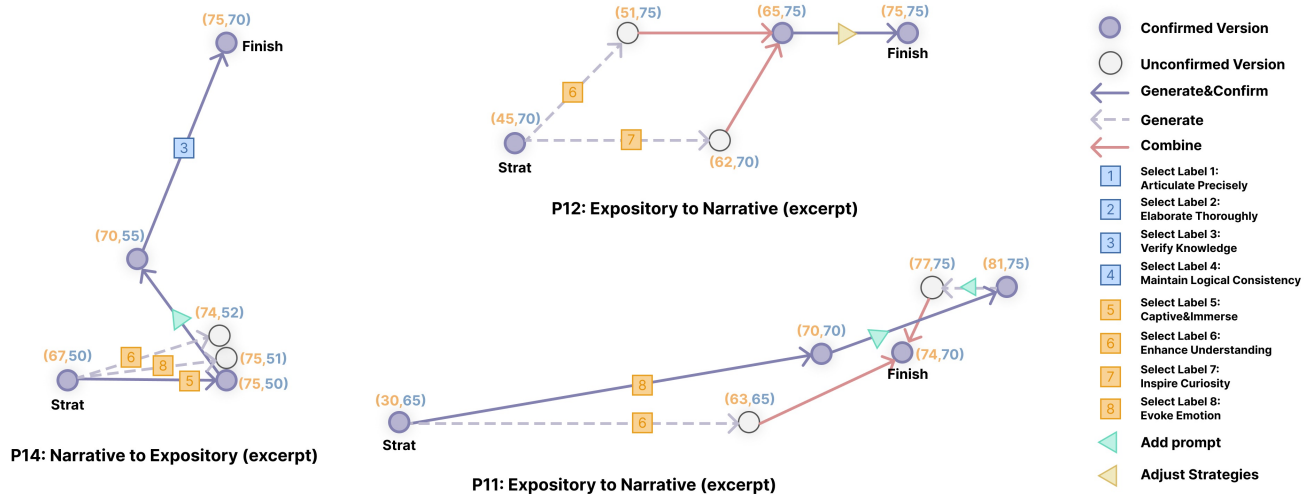


Figure 7: Visualization examples of segment revisions from P11, P12, and P14.

		Spatial Balancing		Baseline		Statistics	
		mean	std	mean	std	p-value	Sig.
NASA-TLX [23]	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 [3]	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	—

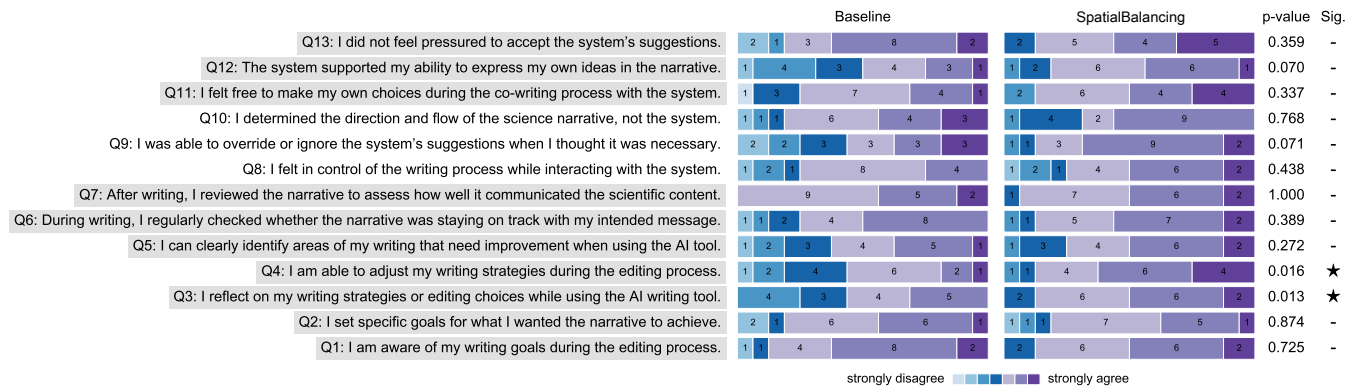
Table 4: The statistical results of NASA-TLX and SUS questionnaires. (\*:  $p < 0.05$  and \*\*:  $p < 0.01$ ).

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

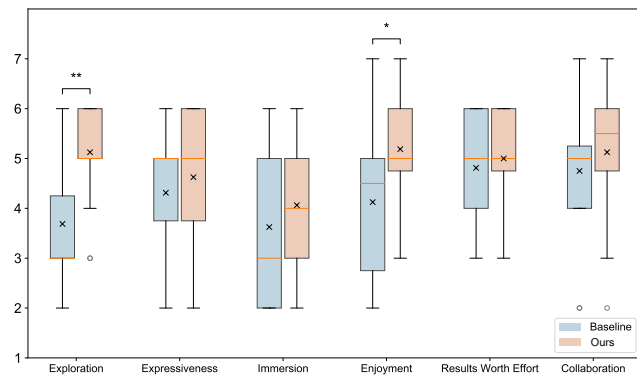
system rather than fixed evaluative benchmarks, using repeated exposure to calibrate their own writing tendencies and reason about audience-specific adjustments.

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



**Figure 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: Q3 ( $M = 5.50$  (Spatial Balancing) vs. 4.63 (Baseline),  $p = .013$ ) and Q4 ( $M = 5.69$  vs. 4.56,  $p = .016$ ); marginal differences in Control: Q9 ( $M = 5.63$  vs. 4.75,  $p = .071$ ) and Autonomy: Q12 ( $M = 5.25$  vs. 4.44,  $p = .070$ ).**



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

once in most cases (Appendix Figure ??). 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 [22]. 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 [53]. 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 [30].

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 [15, 53].

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 [53]. 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 [21, 47], 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.

Our findings also suggest that flexibility should apply not only to how guidance is delivered, but also to how rhetorical goals are represented. While the predefined two-dimensional space served as a useful scaffold in our study, future systems should explore more negotiable goal representations that better reflect writers’ situated and evolving priorities—for example, allowing writers to temporarily swap one rhetorical dimension for a task-specific concern such as tone, audience appropriateness, or platform fit.

**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 [62, 63] 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 Limitations 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 Limits of the Current Baseline and Directions for Future Comparison.** A limitation of the current study is that the baseline captured a realistic editor-and-chat workflow, but did not fully align with Spatial Balancing in how revision support was structured, invoked, and presented during use. In Spatial Balancing, guidance, alternatives, scores, and revision history were integrated into the interface, whereas in the baseline these forms of support were more externally referenced and manually assembled. As a result, the current comparison is better understood as contrasting two different configurations of revision support, rather than as strictly isolating the effect of spatial externalization alone. Accordingly, the present findings should be interpreted as evidence about the benefits and tensions of an integrated revision environment, rather than as a definitive causal test of spatial externalization in isolation. Future studies could isolate the contribution of individual features, such as goal-space visualization, score feedback, and revision traces, through targeted ablation or feature-level comparisons to better identify which aspects of the environment most strongly shaped users’ revision processes.

Although the baseline Excel materials were derived from the same underlying strategy content used in Spatial Balancing, this did not ensure equivalence in how LLM support was enacted during revision. In Spatial Balancing, these supports were embedded into the interaction flow, whereas in the baseline, participants had to manually consult and re-enter them through chat. Future studies should therefore adopt a more harness-equivalent comparison that preserves the same underlying supports while varying only their interface presentation.

**7.2.2 Lack of Outcome-Based Evaluation of Goal Attainment, Text Quality, and Communication Effectiveness.** A key limitation of the current study is that, although we examined how the system supported writers’ goal awareness, strategy reflection, and metacognitive regulation during revision, goal achievement was assessed primarily through participants’ subjective perceptions during the process. We did not independently evaluate whether the final texts ultimately achieved their intended communication goals. Accordingly, our findings should be interpreted as evidence of support for

goal-aware iterative revision, rather than outcome-based evidence of goal attainment. In addition, we did not systematically assess the quality of the resulting texts or their communication effectiveness, such as how they were received by intended audiences or whether they improved reader understanding or engagement. Future work should combine participants' self-assessments with expert review and audience-based evaluations.

**7.2.3 Evaluation Dependency on Proxy Scores.** To demonstrate Spatial Balancing 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.

**7.2.4 Tensions of Externalized Scoring as Guidance.** A further limitation lies not only in what the scores measure, but also in how they guide behavior once externalized in the workspace. In Spatial Balancing, the 2D coordinates were intended as navigational cues for reflection and comparison. However, once these cues become persistent and visually prominent, writers may begin to treat them as goals in themselves. This creates a tension: the same externalized signals that help users stay oriented and compare alternatives may also over-shape revision behavior, encouraging score-seeking, over-trust in system judgments, or reduced attention to situated rhetorical concerns. Future work should therefore make such scoring mechanisms more interpretable and negotiable, for example by exposing the basis of placements, supporting user correction, and allowing writers to define more personalized goal regions or rhetorical dimensions.

**7.2.5 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 reshapes 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.

## Acknowledgments

We thank all participants for taking part in this study, as well as the science communicators who contributed ideas and discussions throughout the project. We also sincerely thank the reviewers for their valuable feedback and suggestions.

## References

- [1] 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.
- [2] Peter Broks. 2006. *Understanding popular science*. McGraw-Hill Education (UK).
- [3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [4] Terry W Burns, D John O'Connor, and Susan M Stockmayer. 2003. Science communication: a contemporary definition. *Public understanding of science* 12, 2 (2003), 183–202. doi:10.1177/09636625030122004
- [5] 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
- [6] Yun Chen, Yiwei Wang, Antoni B. Chan, Jixing Li, and Ray LC. 2025. Once More with (the Right) Feeling: How Historical Fiction Writing Processes of Character Design, Plot Outline, and Context Checking Are Affected by Co-Writing with ChatGPT. In *HCI in Business, Government and Organizations*, Keng Leng Siau and Fiona Fui-Hoon Nah (Eds.). Springer Nature Switzerland, Cham, 79–101. doi:10.1007/978-3-031-92823-9\_7
- [7] 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
- [8] 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
- [9] 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
- [10] 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
- [11] 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
- [12] 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
- [13] Anne DiPardo. 1990. Narrative knowers, expository knowledge: Discourse as a dialectic. *Written communication* 7, 1 (1990), 59–95. doi:10.1177/0741088390007001003
- [14] 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
- [15] 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
- [16] 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
- [17] 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
- [18] 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

- [19] Kexue Fu, Jingfei Huang, Long Ling, Sumin Hong, Yihang Zuo, RAY LC, and Toby Jia-Jun Li. 2026. Vistoria: A Multimodal System to Support Fictional StoryWriting through Instrumental Image-Text Co-Editing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 26. doi:10.1145/3772318.3790400
- [20] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009). doi:10.1515/COMM.2009.026
- [21] Yuanning Han, Ziyi Qiu, Jiale Cheng, and RAY LC. 2024. When Teams Embrace AI: Human Collaboration Strategies in Generative Prompting in a Creative Design Task. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3613904.3642133
- [22] 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
- [23] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adhaporon Khunlerkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IEEE transactions on healthcare systems engineering* 1, 2 (2011), 131–143. doi:10.1016/S0166-4115(08)62386-9
- [24] 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
- [25] 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/SHTT220546
- [26] 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
- [27] 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
- [28] 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
- [29] 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
- [30] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454. doi:10.1007/s00146-010-0272-8
- [31] 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
- [32] 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
- [33] Joe Lambert. 2013. *Digital storytelling: Capturing lives, creating community*. Routledge. doi:10.4324/9780203102329
- [34] Martha Larson, Eamonn Newman, and Gareth JF Jones. 2010. Overview of Video-CLEF 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
- [35] 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.LG] https://arxiv.org/abs/2502.02904
- [36] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsgans, 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, Joon Suk 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
- [37] Norma J. Livo and Sandra A. Rietz. 1986. *Storytelling: Process and Practice*. Libraries Unlimited, Littleton, CO, USA. https://archive.org/details/storytellingproc00livo
- [38] 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
- [39] 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
- [40] Yao Lyu, He Zhang, Shuo Niu, and Jie Cai. 2024. A Preliminary Exploration of YouTube's 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
- [41] 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
- [42] Julia Metag, Florian Winterlin, and Kira Klingner. 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
- [43] 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
- [44] National Academies of Sciences, Engineering, and Medicine. 2017. *Communicating Science Effectively: A Research Agenda*. National Academies Press, Washington, DC, USA. doi:10.17226/23674
- [45] 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
- [46] Fatemeh Rabiee. 2004. Focus-group interview and data analysis. *Proceedings of the nutrition society* 63, 4 (2004), 655–660. doi:10.1079/PNS2004399
- [47] 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
- [48] 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
- [49] 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
- [50] 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
- [51] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary educational psychology* 19, 4 (1994), 460–475. doi:10.1006/ceps.1994.1033
- [52] 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
- [53] 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
- [54] 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
- [55] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminare: 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

- [56] 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
- [57] 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
- [58] Wikipedia contributors. 2024. Popular science – Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Popular\\_science](https://en.wikipedia.org/wiki/Popular_science) Accessed: 2025-03-27.
- [59] 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
- [60] 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
- [61] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and RAY LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the IUI 2022 Workshops: APEX-UI, HAL-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE*. CEUR-WS Team, 56–65. [https://scholars.cityu.edu.hk/en/publications/publication\(d901f5a2-0600-422f-b588-db5a59871961\).html](https://scholars.cityu.edu.hk/en/publications/publication(d901f5a2-0600-422f-b588-db5a59871961).html)
- [62] 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
- [63] 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
- [64] 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
- [65] 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
- [66] 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