Equinox: An LLM-Powered Interface for Visualizing Iterative Revision of Tradeoffs in Science Communication Writing

Anonymous Author(s)

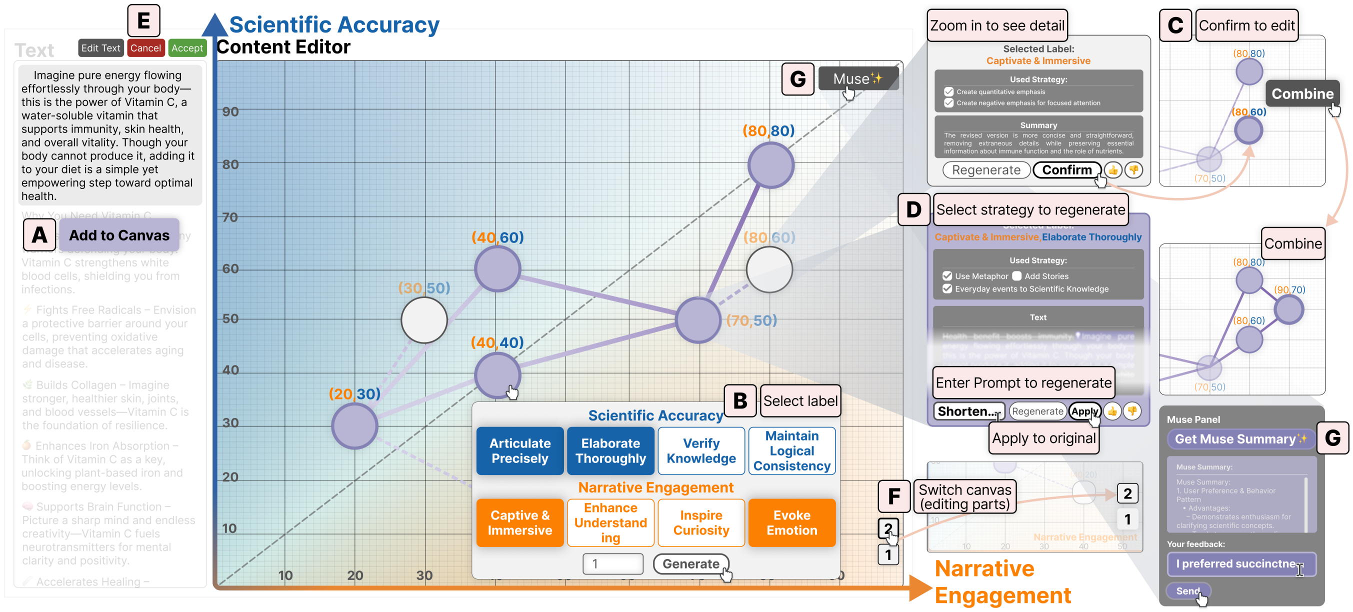

Figure 1: The interaction flow of *Equinox*. A – Selected text segments are plotted as nodes on a 2D space, scored by Narrative Engagement (X-axis) and Scientific Accuracy (Y-axis). B – Users choose from directional labels—four for narrative and four for accuracy—each linked to an LLM-driven strategy. One revision is generated per label and positioned based on its resulting scores. C – Users confirm preferred versions for further refinement; confirmed nodes turn purple. D – Confirmed nodes can be refined via: (1) prompt-based edits, (2) manual strategy adjustments, or (3) combining two versions into a new synthesis. E – Finalized revisions can be applied back into the article for full-context review. F – Additional canvases can be launched to edit other segments independently. G – The Muse module tracks user interaction, offering reflective suggestions and adaptive strategy feedback.

Abstract

Balancing scientific accuracy and narrative engagement is a core challenge in science communication. We present Equinox, a co-writing system that supports revision by visualizing trade-offs in real time via a dual-axis interface. Writers select from strategy-based labels to generate multiple LLM-assisted versions positioned on a coordinate plane reflecting narrative engagement (x-axis) and scientific accuracy (y-axis) score. This layout enables users to compare, refine, and synthesize edits to balance these two dimensions. In a within-subjects study (N=16), Equinox significantly improved metacognitive reflection, flexibility, and creative exploration and enjoyment compared to a baseline. Participants used the coordinate view to surface their communication goals, visually track changes across versions, and make intentional decisions during revision. These findings demonstrate how visualizing revision trade-offs within a structured space enhances writers' strategic awareness and agency, reframing LLM-assisted writing as an intention-driven creative process.

CCS Concepts

• **Human-centered computing → Collaborative interaction**.

Keywords

Narrative Strategy, Science Communication, Mixed-Initiative collaboration, Writing Assistance

UIST '25, September 8 – October 1, 2025, Haeundae, Busan, Korea

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Anonymous Author(s). 2025. Equinox: An LLM-Powered Interface for Visualizing Iterative Revision of Tradeoffs in Science Communication Writing. In *Proceedings of the 2025 ACM Symposium on User Interface Software and*

Technology (UIST '25), September 8 – October 1, 2025, Haeundae, Busan, Korea. ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In the era of social media, online science communication serves as crucial public engagement. The internet democratizes scientific knowledge access while creating challenges—primarily balancing scientific accuracy with narrative engagement during content revision [27, 38, 63].

Large Language Models (LLMs) offer promising support for science communication, as they can synthesize complex information, switch flexibly between tones, and produce stylistic alternatives [15, 40]. These capabilities are particularly valuable when science communication writers often simplify expository knowledge with narratives [55] or embed scientific ideas within storytelling [102]. However, despite these strengths, LLMs offer limited support for more strategic tradeoff demands of science communication writing. Most LLM-powered writing interfaces follow a flat, one-dimensional revision flow [100]: users input a prompt, receive several alternatives, but lack structured guidance on how those revisions affect their communicative goals [34, 83, 91]. There is no mechanism to visualize tradeoffs in the writing process, making it difficult for writers to revise with intention.

Underlying this limitation is a lack of metacognitive support. Metacognition in writing refers to a writer's ability to clarify intentions, monitor progress, and adjust strategies during revision [1, 32]. These demands are especially pronounced in science communication, where writers must balance scientific accuracy with narrative engagement. They must assess how different versions perform along these dimensions, and strategically decide which direction to take next, yet current LLM tools offer little help in managing this complexity. Recent studies have increasingly emphasized these metacognitive challenges in LLM-powered writing [26, 86, 88], underscoring the need for innovative interfaces that help writers track, interpret, and refine their work with greater intentionality.

To address these challenges, we present *Equinox* (**Figure 1**), an interactive interface grounded in metacognitive theory [2, 22, 32, 66, 86]. *Equinox* enables writers to navigate tradeoffs between scientific accuracy and narrative engagement in science writing. The core feature of *Equinox* is a 2D coordinate visualization where each revision is plotted according to its estimated scientific accuracy (Y-axis) and narrative engagement (X-axis) scores. This constant visual feedback allows writers to immediately perceive how different revisions impact their communicative goals, i.e. scientific accuracy versus narrative engagement.

In a within-subjects study with 16 science communicators, *Equinox* significantly outperformed a baseline LLM interface in supporting users' revision processes. Quantitative results showed that *Equinox* increased users' metacognitive reflection and flexibility in using strategies and increased creative exploration. While usability remained comparable, participants reported better idea exploration support when revising with *Equinox*. Qualitatively, Participants found that the coordinate graph externalized abstract writing goals, enabling real-time self-monitoring and strategic planning during revision. By visualizing the tradeoffs between scientific accuracy and narrative engagement, the system supported intentional decision-making and iterative exploration and increased users' confidence in their editorial choices.

This process transformed revision from a fragmented, reactive task into a more coherent and intentional creative workflow. In summary, our contributions include:

- A metacognitively-informed design that operationalizes key cognitive processes—such as intent clarification, monitoring, and strategic flexibility into actionable interaction principles for LLM-assisted science communication writing.
- *Equinox*, an interactive system that instantiates this framework through a 2D coordinate visualization , enabling visual exploration of revision tradeoffs.
- Empirical evidence from a within-subjects study with 16 science communicators showing that *Equinox* improves metacognitive regulation, creative exploration, and writer confidence over a strong LLM baseline.

2 User Scenario: Jenny's Iterative Revision Journey Using *Equinox*

Jenny, a science communicator with a background in immunology, takes pride in her scientific precision. However, she often struggles to make her writing engaging for general audiences. Her latest article on mRNA vaccines, while technically accurate, received editorial feedback as being "too dry" and at risk of losing reader interest. Feeling stuck between preserving rigor and increasing appeal, Jenny turns to *Equinox*.

She begins by dragging a paragraph about how mRNA vaccines stimulate immune responses into the *Equinox* canvas. The system automatically places this segment on the dual-axis plane at (30, 70), confirming her suspicion: the paragraph scores high on Scientific Accuracy but low on Narrative Engagement.

Hovering her cursor over the node reveals eight directional labels—four designed to enhance narrative engagement and four to improve scientific accuracy. Among the engagement-oriented labels, two resonate with Jenny's intent: Evoke Emotion and Inspire Curiosity (**Figure 5**).

She selects both labels simultaneously. *Equinox* generates two new versions of the paragraph, each plotted as a new node: Evoke Emotion: (53, 60), Inspire Curiosity: (40, 70) For Evoke Emotion, the system recommends the strategies "Add Stories" and "Create Negative Emphasis for Focused Attention". This version introduces a brief but vivid story of a retired immunologist who volunteers for an early mRNA vaccine trial, describing her emotional journey of fear and hope—making the science more relatable and emotionally compelling. For Inspire Curiosity, *Equinox* applies the "Question-Answer Hook strategy". The revised paragraph now begins with, "How can a tiny strand of mRNA trigger a full-scale immune defense?" and follows with a mid-paragraph question to deepen reader engagement.

Jenny finds both versions compelling in different ways and uses the Combine feature to merge their strengths. The resulting node, scored at (73, 68), blends the emotive story with a question-driven structure. It begins with a curiosity-sparking question and integrates a personal anecdote that evokes emotional resonance—an effective combination of both narrative strategies (**Figure 6**).

Still, she feels the story could better connect to the underlying science. Using Prompt-based Editing, she enters: "Make the narrative more explicitly linked to immune memory formation." The system generates a new revision with stronger conceptual ties and a clearer explanation of how mRNA vaccines train the immune system. This adjustment improves the scientific accuracy and results in a new node at (73, 77), satisfying both of her communicative goals (**Figure 6**).

Throughout the process, Jenny uses the Zoom-out feature to trace how her ideas have evolved across multiple iterations. She explores branching paths, reflects on each revision's impact, and uses Zoom-in to compare granular content differences between versions (**Figure 3**).

At this point, Jenny activates Muse, the reflective assistant in the canvas corner. Drawing on her revision history—confirmed nodes, label choices, strategies, and prompt edits—Muse spots a trend: she favors emotionally engaging, curiosity-driven edits but hasn't used figurative language. It highlights a literal sentence about immune memory and suggests a metaphor: "Think of mRNA as a 'wanted poster' that trains the immune system's detectives." **(Figure 7)**.

Jenny finds the suggestion intuitive and aligned with her narrative goals. She accepts the recommendation and integrates the metaphor into the paragraph, reinforcing her balance between narrative engagement and scientific accuracy. Finally, Jenny rereads the revised segment in the context of the full article. Satisfied with the improved flow, emotional resonance, and scientific clarity, she proceeds to identify the next section to refine.

3 Related Work

3.1 Science Communication Narrative Design

In the Information Age, online science communication has become increasingly dominant, especially in the popular science field [12, 64]. 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 [10, 47, 67]. Traditionally, science communication content has been categorized into three groups: traditional journalism, live or face-to-face events, and online interactions [9, 10]. 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. Unlike traditional science journalism, which focuses on recent scientific developments and authority, popular science places greater emphasis on entertainment and broadening its scope [7, 23, 96]. 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 [31, 96, 103].

Traditionally, science communication content has been produced by professionals: scientists, journalists, and media makers [10, 25]. However, the rise of online video platforms has democratized content creation, enabling more individuals to produce online popular science content through various digital channels. These include online platforms such as YouTube, social media, blogs, question-and-answer platforms, and podcasts [68, 95, 103]. While this increased accessibility of science content, it also presents a challenge: many of these content creators lack formal training in either science or communication. As a result, the quality of popular sicencecontent can vary significantly [77]. This highlights the need for better guidance and clearer frameworks to support individuals who want to create high-quality online science content.

Science communication narratives are often seen as a delicate balance between two key dimensions: scientific accuracy and narrative engagement [27, 38, 63]. Burns et al. (2003) made a vivid analogy, describing science communication as a form of "mountain climbing," balancing between scientific literacy and science culture [10]. Similarly, Dahlstrom (2014) emphasized that science communication writing inherently involves both narrative and expository elements [19]. Finkler and León (2019) further argued that effective science videos must find a balance between audience engagement and knowledge delivery [31]. In other words, the balance between "narrative engagement" and "scientific accuracy" is a key focus in science communication research [21, 78]. Scholars increasingly emphasize the need to understand how these two dimensions interact in science writing, underscoring the complexity and importance of creating narratives that are compelling and informative [64].

Some scholars have proposed strategies to help creators improve their writing by improving either narrative engagement [19, 30, 35, 38] or scientific precision [49, 54, 69]. Other studies have attempted to explore ways to balance these two elements [5, 55]. However, these studies primarily focus on theoretical contributions and often lack concrete, actionable guidance for content creators.

Recent HCI research has explored how large language models (LLM) can support science communication, with systems focusing on content planning [73, 80], rhetorical enhancement [36, 37, 51], and iterative revision [60, 102]. However, most existing tools focus on either structural planning [80] or localized iterations [51, 52]. There is a general lack of integration between the local edits and the broader narrative design. Moreover, none of these existing tools address the issue of balancing between scientific accuracy and narrative engagement.

This is what *Equinox* attempts to accomplish through its dual-axis diagram design, which provides a visual representation of how individual editing strategies connect to the broader context and goals of the editing process.

3.2 Metacognition in LLM-Powered Writing Tools

Metacognition refers to people's awareness and control over their own cognitive processes [86]. According to Flavell (1979), metacognition comprises both metacognitive knowledge—awareness of one's goals, abilities, and strategies—and metacognitive regulation—the processes by which one plans, monitors, and adapts during cognitive activities [32].

In writing contexts, metacognition refers to the writer's awareness of their cognitive experience and their ability to actively control the processes engaged in the writing task [2]. The application of metacognition in writing encompasses several key processes, including planning, monitoring, revising, summarizing, and evaluating [22]. Having metacognition during writing involves maintaining an ongoing awareness of the creative process to ensure adjustments can be made as needed. Metacognitive skills in writing also involve reflection on one's performance and the ability to correct errors when appropriate [1, 32]. This reflective component allows writers to critically examine their work, identify strengths and weaknesses, and make improvements based on this awareness. This metacognitive approach transforms writers from passive participants to active managers of their own writing process, enhancing both the quality of their written products and their development as writers [89].

However, the introduction of Large Language Models (LLMs) like ChatGPT presents new metacognitive challenges in writing [26, 86, 88]. Unlike traditional writing contexts where cognitive processes may remain implicit, LLM-based writing requires users to externalize goals, formulate effective prompts, and iteratively evaluate and revise system outputs [86]. These processes demand heightened metacognitive monitoring and control, including task decomposition, self-awareness of goals, and well-calibrated confidence [66]. For example, novice users often struggle with prompt formulation because they cannot clearly articulate what they want the system to do [101], and users may misattribute poor outputs either to their own limitations or to model shortcomings without sufficient self-evaluation [53].

To address these challenges, Tankelevitch et al. (2024) propose a dual path framework rooted in metacognitive theory [86]. The first strategy focuses on improving users' metacognitive abilities by providing support mechanisms within LLM interfaces to enhance the process of planning, self-evaluation and self-management. For instance, scaffolding tools that guide users in articulating task goals and decomposing complex writing tasks can directly strengthen self-awareness and planning processes [86]. The second strategy involves reducing the metacognitive demand imposed by LLM systems through thoughtful interface design. There are several existing LLM powered writing tools which are designed to enhance metacognition by providing innovative visual interfaces [17, 18, 46, 62, 81, 84, 92, 105]. Broadly speaking, our system, continuing on this line of research, is based on metacognitive theory and specifically Tankelevitch et al.'s [86] suggestions by improving user's self-awareness and reducing cognitive demands during their writing process through visualization, scaffolding, and real-time feedback. More specifically, our system draws inspiration from Graphologue's [46] node-link diagrams and Polymind's [92] visual diagram approach.

4 Formative Study

4.1 Expert Interview

To better understand the workflows, goals, and tool needs of science communicators, we conducted in-depth interviews with four professionals: a TikTok science animator (20K+ followers), a YouTuber (10K+ subscribers), a science columnist on a Q&A platform (200K+ followers), and an educational video producer. Each interview lasted approximately 90 minutes and focused on three areas: (1) their typical content creation workflow, (2) how they balance communicative goals, and (3) how they use LLM tools in practice. The qualitative findings are as follows:

(1) Balancing Scientific Accuracy and Engagement. Participants described two common workflows in science communication. The knowledge-to-stories approach, favored by those creating platform-independent or long-form content, begins with scientific concepts and adds narrative elements (e.g., examples, metaphors) to enhance engagement. In contrast, the news-to-theories workflow—more typical of real-time or event-driven content—starts with current events or relatable experiences and layers in relevant scientific explanations. Despite differing starting points, all participants emphasized the same challenge: sustaining both scientific rigor and audience interest. One creator noted, "If it's too technical, people stop watching. If it's too entertaining, they call it shallow." Across formats, creators stressed the need to balance clarity, credibility, and emotional connection.

(2) Narrative Strategies and Gaps. To make their writing more engaging, participants reported deliberately applying narrative strategies such as metaphors, real-world analogies, quotations, and personal anecdotes. One creator revised content by adding narrative "hooks" after drafting the science explanation; another explicitly mapped theories to familiar experiences. However, participants also noted that these decisions were largely intuitive and lacked structured support. They expressed a desire for clearer feedback on how well narrative choices served their communicative goals.

(3) LLM Tools: Value and Limitations. All four participants had experimented with LLMs to support writing, primarily for idea generation, tone adjustment, and connecting scientific ideas to familiar concepts. For example, the educator used LLMs to make explanations "more relaxed and child-friendly," while the columnist relied on them to quickly associate trending news with relevant theories. Yet participants also expressed frustration with LLM-generated content—citing issues such as vague language, repetition, lack of specificity, and misalignment with their communicative intent.

These interviews highlight the core challenges of balancing narrative engagement with scientific accuracy, the creative but undersupported role of narrative strategies, and the untapped potential of LLM tools in this domain.

4.2 Design Space for Science Communication Narrative Design

Based on the results from the pilot interviews, 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 accuracy. We searched keywords "science communication" OR "scientific writing" OR "popular science" AND "strategy" OR "strategies" OR "method" in Google Scholar, the ACM Digital Library, and the IEEE Xplore Digital Library. After screening the abstract and full paper, we selected 47 papers, across Education (N=5), Psychology (N=7), Communication Studies (N=27), Linguistics and Writing (N=4), and HCI (N=6). We identified a total of 25 strategies from these selected papers. By using open coding [42] and design space analysis [13] methods, two authors developed and organized a design space (**Table 3**).

In this design space, we categorized the 25 identified strategies into three groups: those that enhance narrative engagement (N=10),

Table 1: Labels of Science Communication Writing Strategies.

Scientific Accuracy			
Label 1 **Articulate Precisely**	**Label 2** **Elaborate Thoroughly**	**Label 3** **Verify Knowledge**	**Label 4** **Maintain Logical Consistency**
Communicates scientific concepts with accuracy and clarity, using appropriate terminology and well-defined language to prevent ambiguity or misinterpretation [45, 49, 65].	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 [55, 69].	Supports claims with credible sources, data, or reasoning, allowing audiences to feel more trustworthy of the given information [55, 74].	Ensures that arguments and explanations are coherent and internally consistent, following a clear logical structure [90].
Strategies: (4) Acknowledge Uncertainties, (5) Consistent Terminology, (18) Simplify and abstract language, (19) Clarify Key Terms, (21) Repeat key point(s) or question(s), (22) Emphasize with Numbers	**Strategies:** (3) Step-by-Step Explanation, (4) Acknowledge Uncertainties, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (25) Tie Science to Current Events	**Strategies:** (2) Rigorous Source Verification, (6) Citations & Quotes, (7) Everyday Events to Scientific Insights, (22) Emphasize with Numbers, (25) Tie Science to Current Events	**Strategies:** (1) Layered Transitions, (3) Step-by-Step Explanation, (20) Key Point Recap, (23) Strengthen the Connections Between Content
Narrative Engagement			
Label 5 **Captivate & Immerse**	**Label 6** **Enhance Understanding**	**Label 7** **Inspire Curiosity**	**Label 8** **Evoke Emotion**
Engages the audience's attention and draws them into the narrative or content flow by adding stories [38, 59] or using intriguing language [30, 65].	Help audiences to grasp complex scientific ideas using rational, structural content or vivid analogies, visualizations [30, 38, 43].	Stimulates the audience's desire to learn more and have motivation to further explore by applying different forms of questions [56].	Creates an emotional response, positive or negative, and makes the audience feel connected to the content, even immerse themselves in the described scenario [38, 75].
Strategies: (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal, (11) Use metaphors, (12) Inject humor, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire	**Strategies:** (11) Use metaphors, (13) Add real-world supporting examples, (14) Add stories, (15) Add an imagery description, (21) Repeat key point(s) or question(s), (23) Strengthen the Connections Between Content, (24) Present Balanced Views, (25) Tie Science to Current Events	**Strategies:** (8) Question-Answer Hook, (9) Reflection Question, (10) Suspense-Driven Reveal	**Strategies:** (9) Reflection Question, (12) Inject humor, (14) Add stories, (16) Create negative emphasis for focused attention, (17) Make positive emotion to expand action repertoire, (21) Repeat key point(s) or question(s)

Note. **Specific information about each strategy (e.g., definitions, examples) is presented in Table 3.**

those that enhance scientific accuracy (N=7), and those that enhance both (N=8).

Then, we conducted a Focus Group Discussion (FGD) [72] with the four experts we had previously interviewed. They all affirmed the accuracy of our design space, specifically the strategies and their categorization. Additionally, the experts suggested labeling the strategies based on their effects on science communication writing. In this way, the labels highlight the effects of each strategy, helping to build a clearer and more structured framework that makes the design space more comprehensive. Furthermore, the experts emphasized that establishing these labels is crucial for making the design space more easily usable for the users. We agreed that establishing these labels provides a systematic way to categorize the effect of different writing strategies on science communication writing. Therefore, based on the results of the FGD, we established eight labels in total (**Table 1**).

4.3 Design Goals

Drawing from the findings of the formative study and existing literature on science communication and metacognition, we have established the following design goals:

DG1. Visualize Trade-offs to Ease Balancing Effort. As previous literatures highlight the importance of balancing science accuracy and narrative engagement in science communication writing [27, 38, 63], and our formative interviews (See Section 4.1) show that creators grapple with delivering both accurate content and engaging storytelling. Meanwhile, recent research on LLM highlights users must maintain a well-adjusted level of confidence in their own ability to evaluate this output and not blindly accept generated content [86]. Consequently, the system should make these dual goals visible and less mentally taxing to balance between scientific accuracy and narrative engagement, thereby helping creators maintain clarity of purpose during the writing process without cognitive overload.

DG2. Guide Revisions with Strategy Scaffolds to Balance Tradeoffs. Prior literature documents many techniques to address distinct communication objectives (See Section 4.2). Yet, LLM usage requires explicit task decomposition and self-directed prompting, which demand metacognitive control [86]. The system should therefore scaffold trategies—offering prompts, labels, etc. that help users systematically select and apply approaches best suited to their communication goals. This reduces the burden of recalling strategies and allows for more deliberate, goal-oriented writing process.

DG3. Enable Flexible Exploration Through Multi-Version Revision. Effective writing often emerges through multiple drafting cycles and iterative refinement [28], and these needs become even more pronounced in LLM supported writing—where prompt specificity and the inherent variability of LLM outputs make it essential to explore and synthesize multiple solutions while keeping one's overarching goals in view [50, 86]. Because each LLM iteration may produce new or unexpected ideas, creators must remain flexible in revisiting earlier revisions, combining promising elements, or reverting to a previous version if it better supports their broader communicative aims [34, 91]. Hence, the system should enable users to generate, compare, and merge multiple versions. By offering non-linear history tracking and granular editing controls, creators can reinforcing their metacognitive reflection and flexibility. This approach also has the potential to foster creativity by encouraging experimentation and the discovery of unconventional approaches.

DG4. Embed Reflection Within Iteration to Support Self-Monitoring. Effective science communication writing with LLMs involves not only generating content, but also navigating iterative cycles of revision and evaluation [51, 60]. During these cycles, writers must continuously monitor progress toward communicative goals and adjust based on their evolving intent. However, in everyday interactions with LLMs, such self-monitoring is often missing or implicit [58]. Metacognitive theory emphasizes that monitoring—assessing alignment between current output and original goals, detecting overreliance on familiar strategies, or noticing when a revision veers off-course—is central to effective regulation [86]. To support metacognitive monitoring, the system should embed reflective signals directly within the revision workflow—e.g., through visual cues or checkpoints—that surface self-assessment opportunities and make reflection a natural part of revision.

5 System Design and Implementation

5.1 Interface & Features

Equinox features a text editor (left) and an exploratory canvas (right) **(Figure 2)**. Users can add selected text—ranging from a full article to a paragraph or sentence into the canvas for iterative revision, where each version is plotted along two axes: Narrative Engagement (x-axis) and Scientific Accuracy (y-axis). Gray dots represent exploratory drafts, while purple dots indicate confirmed user selections, which can be further revised by reselecting labels or fine-tuning the content. This visualization makes the revision process and decision points transparent, supporting users in balancing scientific accuracy and narrative engagement throughout iterative editing. The following sections introduce each feature in detail.

As shown in **Figure 3**. The canvas interface of the *Equinox* provides three levels of visual representation for the multiple versions generated during the tradeoff iteration process. When zoomed out (0–30%), each version is shown as a simple point, offering an overview of the iteration landscape. At medium zoom levels (40–70%), users can view a summary of changes for each version compared to its predecessor, along with the selected label and strategy. In full zoom (80–100%), detailed content changes are displayed, with differences from the original text highlighted for clarity.

Real-time Two-Axis Feedback (DG1& DG4). (**Figure 4**) Leveraging insights from metacognitive research [86], authors benefit from explicit feedback that reduces the cognitive burden of juggling multiple objectives (DG1) and allows self-monitoring of revision progress and alignment with writing intention(DG4). In *Equinox*, each version of the text is plotted as a point in a two-dimensional space, with one axis representing *narrative engagement* and the other *scientific accuracy*. A "Scorer Agent," trained on audience ratings, assigns scores whenever users drag a new piece of text into the canvas to create a node or perform additional edits that generate additional nodes. These scores determine the position of each node on the coordinate axes. This immediate visualization helps creators monitor their balance between scientific accuracy and narrative engagement, enabling them to maintain clarity of purpose and goals in writing.

Strategy Recommendation via Eight Labels (DG1 & DG2). (**Figure 5**) Science communication research highlights numerous narrative and explanatory strategies, but introducing all of them at once can overwhelm users. Instead, *Equinox* offers an eight-label taxonomy (e.g., *inspire curiosity* or *elaborate thoroughly* to represent core revision goals. These labels are informed by the formative study expert interview and literature review (See **Section 4**). Four of these labels focus on improving scientific accuracy, while the other four encourage improvements in narrative engagement (**Table 1**). When a node is confirmed, the system invites the user to apply one or more labels, spawning new versions that emphasize these chosen strategies. Presenting these targeted options scaffolds the revision process (DG2) and lessens mental load (DG1), because users can systematically select a path that leads to either "scientific accuracy" or "narrative engagement" without needing to recall every possibility themselves. This eight-label structure also offers a revision framework, reminding creators of directions they might not have initially considered. By doing so, they exercise metacognitive control, deliberately steering each iteration toward their intended revision direction.

Fine-Grained Control for Specific Versions (DG3). (**Figure 6**) After exploring different branches, users can refine a single node in greater depth. After a user confirms a bottom, its color changes to purple, and three fine-tuning operations become available. The other unconfirmed points remain gray, allowing the user to clearly distinguish between confirmed and unconfirmed nodes through their visual connections. Three possible refinements include toggling previously applied strategies, providing customized prompts (e.g., "try a different metaphor" or "make this more concise"), or merging two versions to preserve strong elements from each. These fine-grained actions underscore metacognitive flexibility: creators can adapt or pivot without discarding prior work. Aligned with DG3, this feature facilitates iterative metacognitive regulation by enabling cycles of exploration (metacognitive flexibility) and synthesis (monitoring and control), allowing users to evaluate competing versions and consolidate revisions in alignment with their communicative goals.

Tree-Based Content Generation (DG3). (**Figure 4**) When text is dragged into the exploratory canvas, it becomes a root node. From

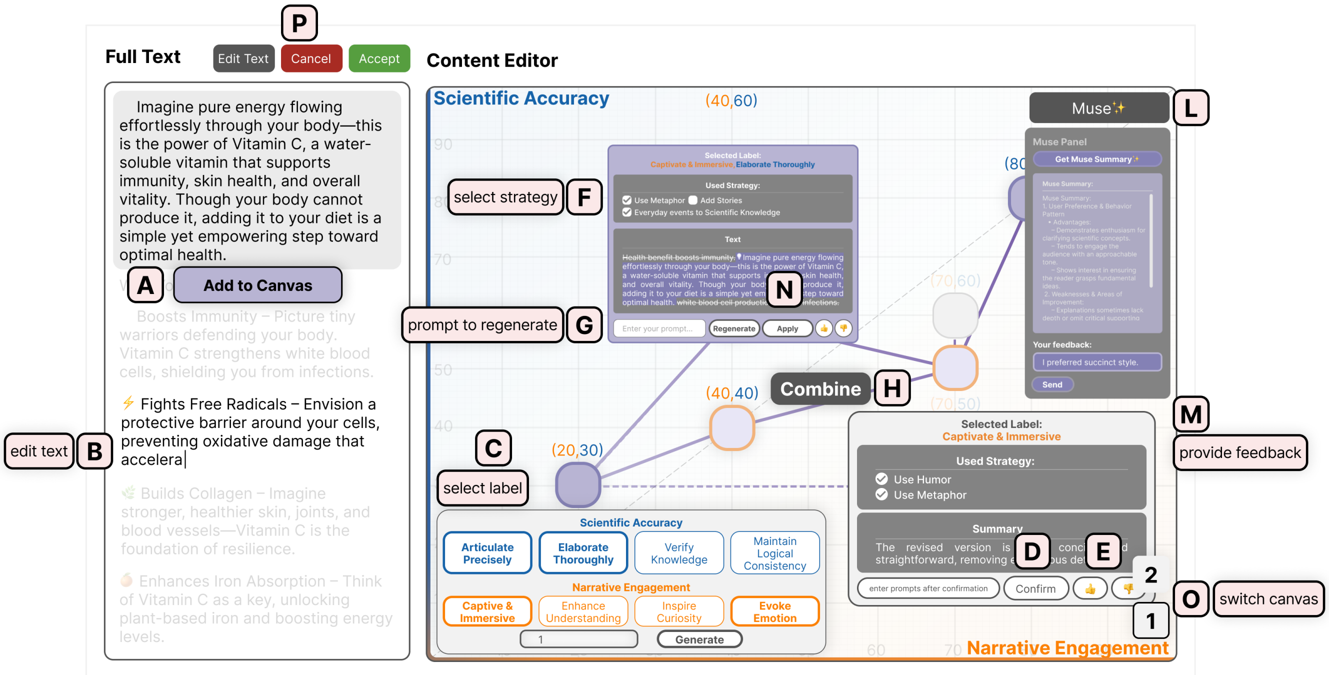

Figure 2: The *Equinox* interface has two main sections: a text editor on the left for placing and directly editing source text (B), and a canvas on the right for revising selected segments (A). In the center, a visualization tracks iteration scores across narrative engagement and scientific accuracy 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).

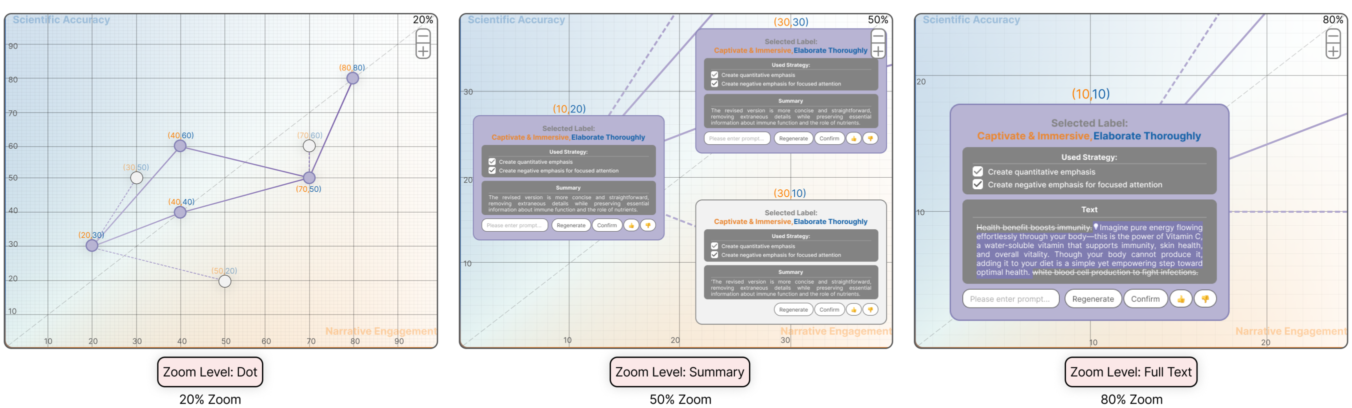

Figure 3: *Equinox* canvas supports three zoom levels: dots for version overview (0–30%), change summaries with labels and strategies (40–70%), and full content with highlights of edits (80–100%).

there, diverse branches (child nodes) emerge as users select different labels or customized instructions. The resulting tree visually traces each exploration path, revealing varying levels of detail from simple node icons to full-text displays. This structure fosters reflection on decision-making and encourages the comparison of multiple alternatives, echoing DG4's principle of iterative, multi-version revision. It also bolsters metacognitive flexibility, enabling creators to identify promising branches, revert to earlier nodes when beneficial, and continue refining or restructuring the content.

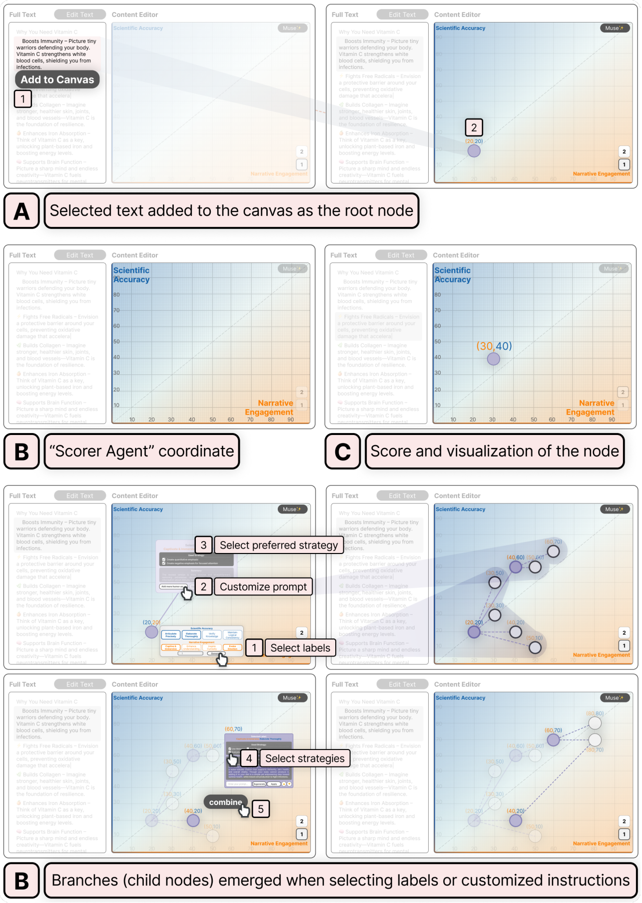

Figure 4: Visualizing the revision process of balancing scientific accuracy and narrative engagement: each branch in the version tree reflects iterations driven by different selected labels and customized instructions. All versions are recorded and can be revisited or further edited.

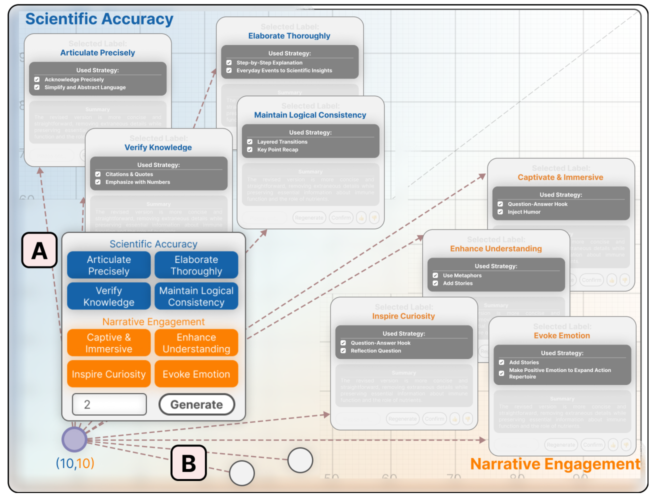

Figure 5: Among the eight revision labels provided in *Equinox*, four are designed to enhance narrative engagement, while the other four focus on improving scientific accuracy. Users can select one or more labels and specify how many versions they want to generate under each. The system then produces label-guided revisions using the LLM, and visualizes the results as points on a 2D coordinate plane in real time, enabling users to see how different strategies shift the text's position along the narrative engagement and scientific accuracy tradeoff space.

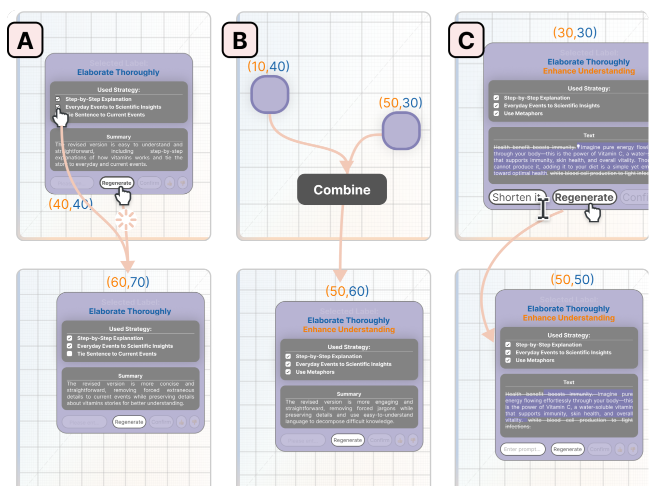

Figure 6: After generating content based on selected labels, users can fine-tune the resulting nodes in several ways: A – Modify the strategy list used by the recommender agent for a specific label. B – Combine different nodes to apply strategies from two labels simultaneously. C – Input custom prompts to refine the current node with personalized edits.

"Muse" Reflective Feedback (DG3& DG4). **(Figure 7)** A "Muse" agent continuously monitors user behavior—such as node confirmations, upvotes or downvotes on single nodes to convey AI strategy preferences, strategy selections, and choices regarding scientific accuracy versus narrative engagement—and synthesizes these inputs into actionable feedback. Building on findings that guided reflection enhances metacognitive skill [86], Muse highlights patterns in the user's editing process. Specifically, Muse presents feedback in a structured format, organized into sections such as strengths, weaknesses, user patterns & goals (whether they successfully balance scientific accuracy and narrative engagement or over-rely on certain strategies), and strategy suggestions to the current content. This structured presentation offers users a clear channel for reflecting on their revision process. After receiving feedback from Muse, users can respond by indicating whether they accept or reject the suggestions. This feedback is then passed to the Recommender Agent, allowing the system to refine future strategy recommendations accordingly., Muse supports greater self-awareness (DG4) and encourages iterative refinement (DG3). As users adjust their approach,*Equinox* adapts accordingly, refining its recommendations to match evolving intentions and individual styles.

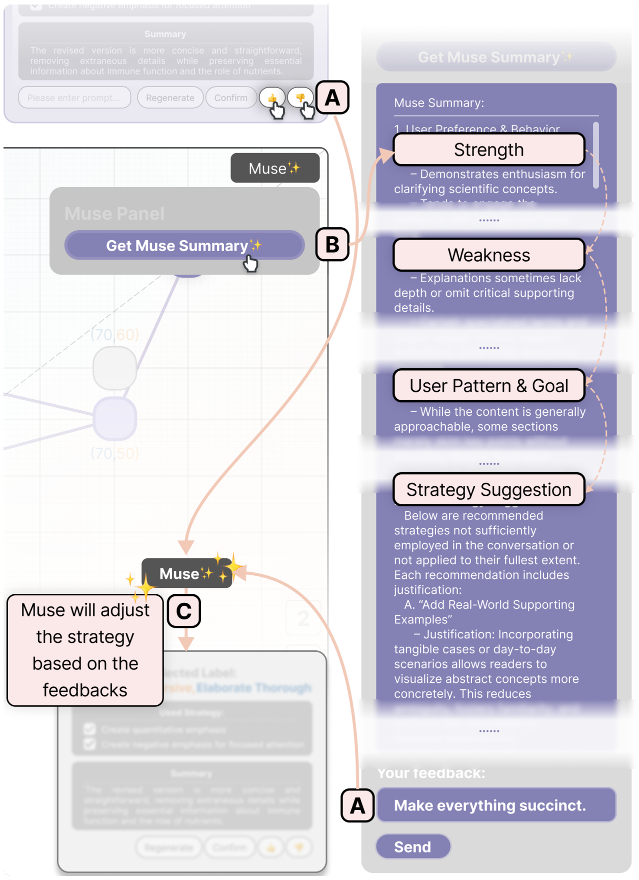

Figure 7: Users can click the bottom button (B) to receive feedback from Muse on their current iterative revision process. They can also like or dislike individual nodes (A). Muse's feedback includes four components: strengths, weaknesses, user patterns & goals, and strategy suggestions for the current version. Users may respond to the feedback, and their input will inform future strategy recommendations to better align with their preferences (C).

5.2 *Equinox* Backend and Implementation

The backend of *Equinox* comprises several LLM-based agents organized into two main modules: a generation module and a reinforcement module. The overall pipeline is in **Figure 8**.

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

Recommender Agent: The recommender agent's core function is to generate multiple strategy combinations based on a user-selected goal. When a user chooses a label, the agent analyzes the current textual features to identify the best combination from its associated strategy set **(Table 1)**. Prompts are constructed using in-context learning and chain-of-thought principles. The agent considers several factors when recommending strategies for each label, including strategy definitions, usage guides, examples, and the original text's role within the broader context of the entire text to recommend the most suitable strategies. The final output consists of multiple strategy combinations, which are then passed to the scorer to filter and select the top-scoring versions that best meet the requirements.

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

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

To support this, we curated a high-quality dataset of 45 science texts from five domains, varying in length and narrative style. Each text was revised by a science communication expert and annotated by 27 non-experts using a rubric developed by three domain experts. The rubric incorporated sub-dimensions of narrative engagement [11] and scientific accuracy [19], emphasizing perceived credibility over strict factual correctness. Scores were normalized to a 0–100 scale and used to fine-tune a GPT-4o model via a small-sample learning strategy[1], enabling it to approximate human evaluative behavior across both axes. The scorer agent is powered by this fine-tuned model. Scoring prompts are consistent with those used during fine-tuning. Details on dataset construction and model training are provided in **Appendix A.2**.

Survey results **Figure 9** reveal clear trade-offs between the two axes: story-like texts scored higher on narrative engagement but lower on scientific accuracy, while expository texts showed the reverse. Infotainment-style texts typically balanced both. Longer texts consistently outperformed shorter ones, likely due to richer explanations and narrative depth. These findings demonstrate the scorer's effectiveness in capturing nuanced audience preferences and guiding users toward more balanced revisions.

To ensure consistent and reliable evaluation, we adopt a **comparative scoring strategy**, aligned with findings from comparative judgement research [70]. Rather than evaluating revisions in isolation, the agent receives the original version and its score as historical context, enabling more accurate scoring that reflects revision trajectories.

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

5.2.2 Reinforcement Module Since user iterations form a tree of nodes enriched with valuable data (selected labels, prompts, likes

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

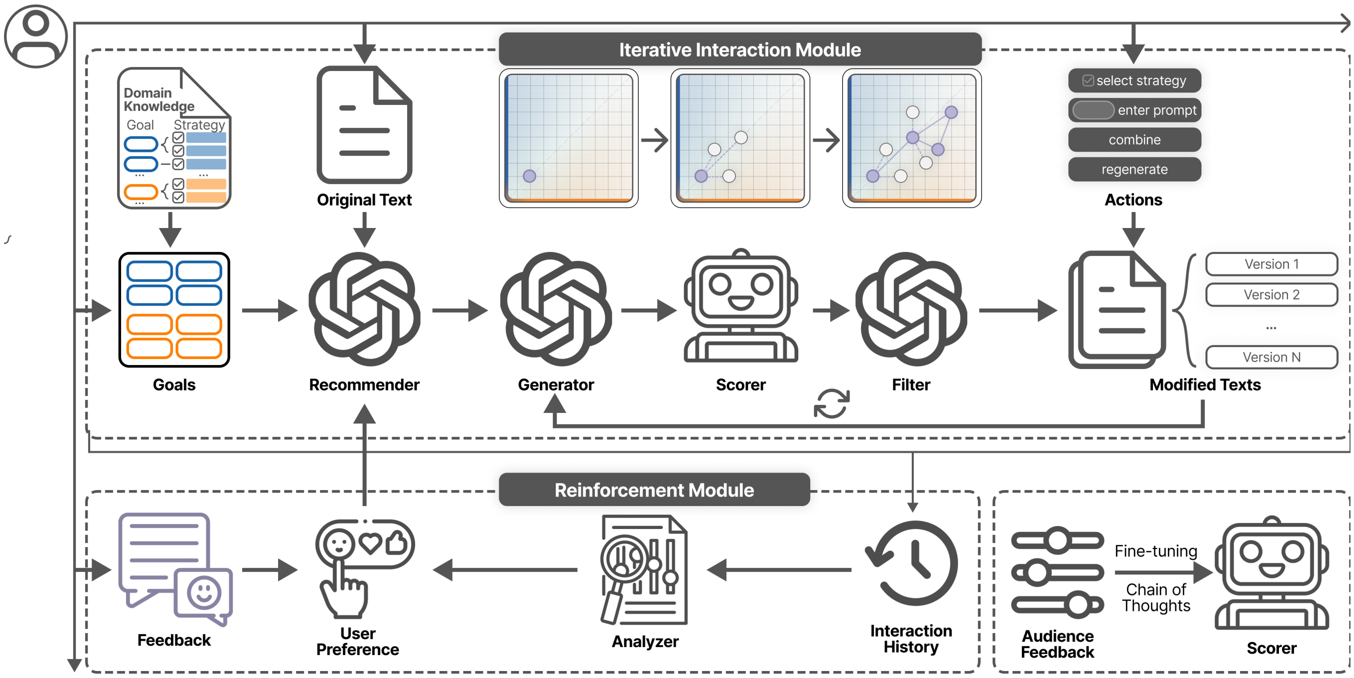

Figure 8: *Equinox* backend overview. *Equinox* consists of two core modules: (1) The Iterative Interaction Module, where LLM-based agents—Recommender, Generator, Scorer, and Filter—collaboratively produce and evaluate multiple content versions based on narrative engagement and scientific accuracy; and (2) the Reinforcement Module, which captures user feedback and inference based on interaction history of user behaviors to refine strategy recommendations through the Analyzer agent. This architecture supports adaptive text revision.

/dislikes, and feedback), we developed an analyzer agent to harness both the explicit and implicit signals from these interactions. The analyzer agent captures behavioral data during the iterative process and uses chain-of-thought prompts to interpret user revision behavior.

Analyzer Agent: The analysis focuses on two goals: (1) detecting common editing patterns, such as stylistic preferences, trade-offs between scientific accuracy and narrative engagement, and user strengths or weaknesses; and (2) surfacing alternatives or underused strategy directions. These insights are passed to the Muse component (Section 5.1). After getting the feedback, another function will update the analysis of Analyzer Agent with real-time user feedback (e.g., suggestion approvals or continued edits) to the Recommender Agent to refine strategy recommendations, guiding the next iteration toward better alignment with user preferences. The feedback loop ensures the system adapts continuously to the user's evolving goals and better balances narrative engagement and scientific accuracy in science communication writing during the revision process.

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

For the AI agents, we employ different LLMs tailored to their functional roles. The recommender, generator, and filter agents are powered by the GPT-4o-mini model, optimized for fast, high-quality content generation. The analyzer agent, which requires deeper reasoning to interpret user behavior and editing patterns, is supported by the GPT-o1 model—a reasoning-oriented LLM. For the scorer agent, it is powered by a fine-tuned GPT-4o model using a small-sample learning strategy[4]. The frontend into predefined prompt templates and communicates with the remote LLMs to obtain results. This modular design allows us to tailor agent behavior based on context while maintaining flexibility in prompt construction and LLM selection. The detailed use of prompts in the backend can be found in the **Appendix A.7**.

6 User Study

To further understand the effect of the *Equinox* system on users' experience during the science communication narrative writing process—particularly its impact on users' cognition and human-AI collaboration behavior patterns—we conducted a within-subjects user study involving 16 participants with prior experience in science communication. All participants were recruited from a local university. Each participant completed four text editing tasks: two using the *Equinox* system and two using a baseline system.

[2]https://flask.palletsprojects.com/en/stable/
[3]https://github.com/wbkd/react-flow/
[4]https://platform.openai.com/docs/guides/fine-tuning?utm_source=chatgpt.com

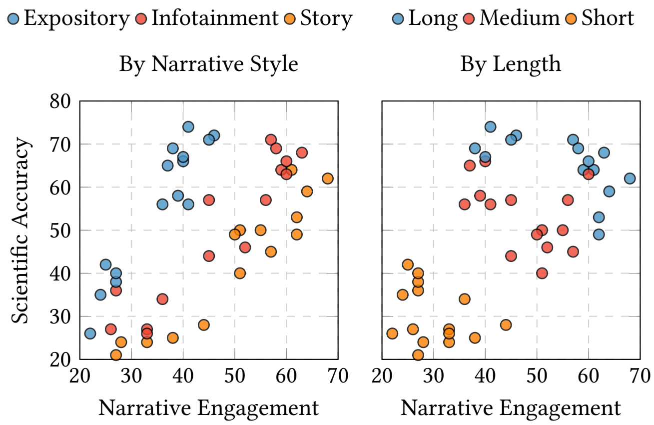

Figure 9: Each point represents one of 45 science communication texts, plotted by its average audience rating for narrative engagement (x-axis) and scientific accuracy (y-axis), based on 27 crowd-sourced rubric-based evaluations per text. The left panel groups texts by narrative style: Expository (informational, fact-focused), Story (highly narrative), and infotainment (represents infotainment-style revisions that blend factual exposition with narrative strategies). The right panel groups texts by length (Short=50 words, Medium=150 words, Long=300 words).

The baseline system used in this study was Cursor, which integrates GPT-4o as its backend. Cursor was selected due to its support for targeted text modifications. In both conditions, participants were provided with an Excel file containing a comprehensive strategy table. This table included the strategy name, definition, usage instructions, examples, and corresponding labels. Participants were encouraged to use this table as a reference and to copy-paste content into the prompt area as needed during the tasks.

6.1 Participants

We recruited 16 participants (9 male, 7 female) aged between 24 and 31 (M = 26.9, SD = 2.0), all of whom held postgraduate degrees or higher. Most were PhD students, postdoctoral researchers, or university faculty members affiliated with a local university, possessing substantial experience in academic work, teaching, or public science communication.

13 participants reported hands-on experience creating science communication content, including teaching undergraduate courses, producing explanatory videos, and translating complex scientific ideas for general audiences. Six of the 13 participants held hybrid professional roles that extended beyond academia, such as science content creators, media producers, journalists, or educators. The remaining three primarily identified as consumers of science communication media.

Regarding their use of AI writing tools, six participants reported daily use, six reported weekly use, and four used them occasionally. In terms of writing confidence, half of the participants (n = 8) self-identified as confident writers, indicating a strong belief in their ability to convey scientific information clearly and persuasively. The remaining half (n = 8) reported a neutral stance, reflecting a moderate level of self-assurance and a potential openness to support or improvement in articulating complex concepts for diverse audiences. The demographic information of these participants are in **Appendix A.5**.

6.2 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 *Equinox* 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:

For the expository text: "I have a scientific narratives. How can I make it more engaging and interesting for an online science video?"

For the narrative storytelling text: "I have a story as online science video narratives. How can I link it with more scientific concepts and add scientific credibility?"

The length of each text averaged 297.75 words (SD = 19.64). The complete versions of the source texts used for the editing tasks are provided in **Appendix A.3**.

To ensure balanced exposure and mitigate order effects or personal topic preferences, we counterbalanced both the system order (*Equinox* 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 *Equinox* condition.

6.3 Post-Task Survey and Instruments

After completing both conditions, participants filled out a post-task survey including standardized instruments: the System Usability Scale (SUS)[8], NASA-TLX for workload[41], and the Creative Self-Efficacy Index (CSI) [16], with one item adapted to: "I think this system supported me in developing ideas or text collaboratively."

We also developed a concise co-creation survey focused on two metacognitive constructs drawn from cognitive psychology [32, 79]. Metacognitive knowledge assesses users' awareness of cognitive goals (e.g., "I am aware of my writing goals during the editing process"). Metacognitive regulation captured planning, monitoring, and evaluation [71] (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 [79] and aligned with recent insights on AI-induced metacognitive demands [86].

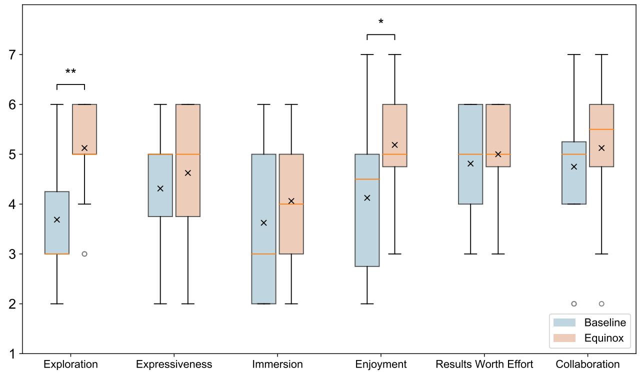

Figure 10: The results of CSI questionnaire. ($*$: $p < 0.05$ and $**$: $p < 0.01$). Participants rated *Equinox* significantly higher in terms of "Exploration" (M = 5.13 (*Equinox*) vs. 3.69 (Baseline), p = .004) and "Enjoyment" (M = 5.19 vs. 4.13, p = .039)

To assess perceived control during co-creation, we included a brief set of items inspired by Human-AI interaction principles [93], evaluating participants' influence over system outputs and narrative direction. We also assessed perceived autonomy based on Self-Determination Theory [24], focusing on decision-making freedom, expressive latitude, and resistance to system pressure. The full list of items related to metacognition, perceived control, and perceived autonomy can be found in **Appendix A.4**.

Finally, participants evaluated their own edits with two targeted questions: "For the expository text, to what extent do you think you improved its narrative engagement?" and "For the narrative text, to what extent do you think you improved its scientific accuracy?" All items of NASA-TLX, SUS, CSI, co-creation survey and these two target questions used a 7-point Likert scale. Following task completion, each participant joined a 15-minute semi-structured interview designed to capture deeper qualitative insights into their cognitive processes, feature usage, perceived system value, and moments of creative difficulty or breakthrough. This interview complemented survey responses and enriched our understanding of user experience across both system conditions.

7 Results

We present analysis of survey responses, participants' interactions with *Equinox* (e.g., how they explored node-based revisions and interacted with strategy labels), observations, and interviews. This section describes participants' overall assessment of *Equinox*, their workflows, and how the system supported metacognition through visual interaction and iterative co-editing during the writing process.

We began by evaluating participants' cognitive workload and perception of usability using the NASA-TLX and SUS questionnaires (**Table 2**). NASA-TLX results showed no significant differences between *Equinox* and the baseline across all six dimensions. This indicates that *Equinox*, despite its expanded feature set, does not impose additional cognitive burden on users. Regarding system usability, the SUS results revealed two statistically significant differences. *Equinox* was perceived as significantly more functionally integrated (Q5, p = .003), but also as requiring more user support (Q4, p = .031). These results suggest that while the system offers richer and more sophisticated capabilities, it also introduces a learning curve, particularly for first-time users. Nevertheless, the overall usability scores were comparable between*Equinox* (M = 70.78) and the baseline (M = 68.44), indicating that both systems were generally regarded as usable. To further assess the creative support provided by the system, we administered the CSI questionnaire. Participants rated *Equinox* significantly higher in terms of "Exploration" (M = 5.13 (*Equinox*) vs. 3.69 (Baseline), p = .004) and "Enjoyment" (M = 5.19 vs. 4.13, p = .039), suggesting that the system better supported users in exploring diverse narrative directions and made the writing experience more enjoyable. *Equinox* showed slightly higher averages across all items in CSI. These results indicate that *Equinox* effectively fosters idea exploration and engagement, key factors in creative writing, without sacrificing usability or increasing user burden. Ths results of CSI is in **Figure 10**.

To evaluate the system's impact on users' metacognitive processes, we measured metacognitive knowledge and regulation of participants using *Equinox* to revision two articles from two directions. *Equinox* received significantly higher ratings than the baseline on two dimensions of metacognitive regulation: RQ3- reflecting on one's own strategies (M = 5.50 vs. 4.63, p = .013) and RQ4- adjusting strategies during the editing process (M = 5.69 vs. 4.56, p = .016). These results suggest that *Equinox* supports users in dynamically managing their writing strategies. For other dimensions, such as identifying areas for improvement, goal setting, and progress monitoring, *Equinox* also showed higher means.

In terms of perceived control and autonomy, participants rated *Equinox* slightly higher across all items, especially in their ability to RQ9- override system suggestions (M = 5.63 vs. 4.75, p = .071) and RQ12- express their own ideas (M = 5.25 vs. 4.44, p = .070), although these did not reach significance. These trends indicate that *Equinox* fosters a stronger sense of authorship and agency in the LLM-supported writing process.

The results of metacognition, control and autonomy are shown in **Figure 11**.

As shown in **Figure 12**, participants generally found the Real-time two-axis feedback function most useful (M = 5.94, SD = 1.18), followed closely by Eight labels to choose directions (M = 5.81, SD = 1.17) and Re-generate with customized prompts (M = 5.88, SD = 0.81). These features were particularly appreciated for providing guidance and support during the creative process. While all functions received relatively positive ratings (above 4.5 on average), these results suggest that our system is effective and provides meaningful support for users' creative workflows.

The quantitative data suggests that *Equinox* system effectively enhances metacognitive abilities and facilitates users' creative thinking through iteration on balancing between scientific accuracy and narrative engagement. In the following sections, we provide explanations based on user interaction data, qualitative feedback and observations during the editing process to illustrate how*Equinox*'s design features, especially the 2D coordinate visualization and scaffolding labels enhance metacognition and creativity in practice.

		Equinox		Baseline		Statistics	
		mean	std	mean	std	p-value	Sig.
NASA-TLX [41]	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 [8]	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 2: The statistic results of NASA-TLX and SUS questionnaires. ($*$: $p < 0.05$ and $**$: $p < 0.01$).

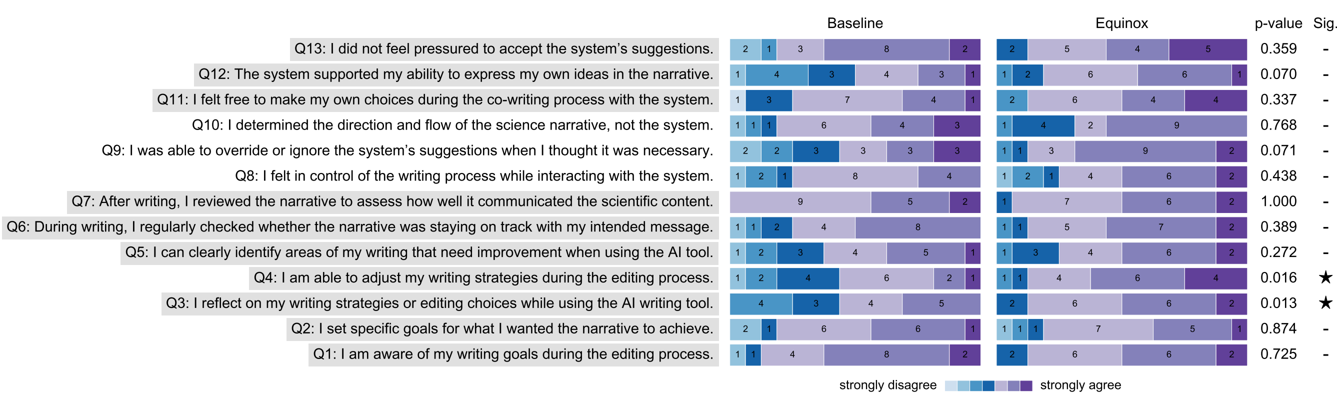

Figure 11: Results of the Metacognition (Q1–Q7), Control (Q8–Q10), and Autonomy (Q11–Q13) questionnaires (p < .05 marked with *; p < .01 with **). Significant differences were observed in Metacognition: RQ3 (M = 5.50 (*Equinox*) vs. 4.63 (Baseline), p = .013) and RQ4 (M = 5.69 vs. 4.56, p = .016); marginal differences in Control: RQ9 (M = 5.63 vs. 4.75, p = .071) and Autonomy: RQ12 (M = 5.25 vs. 4.44, p = .070).

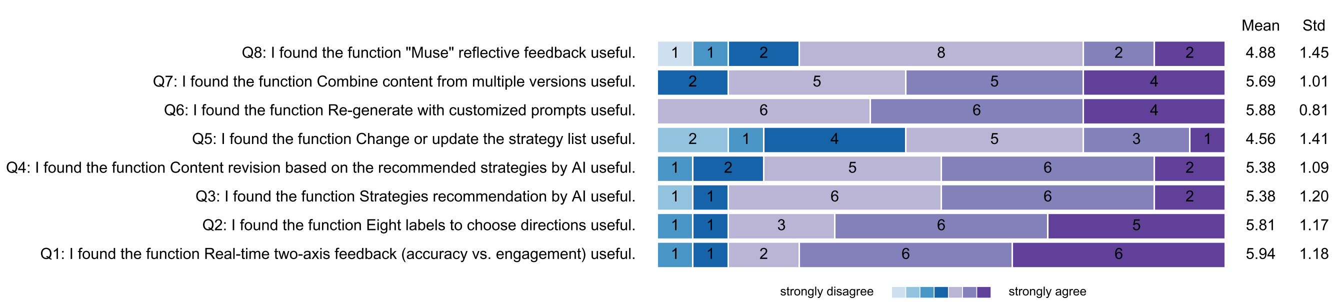

Figure 12: Functional evaulation of *Equinox*.

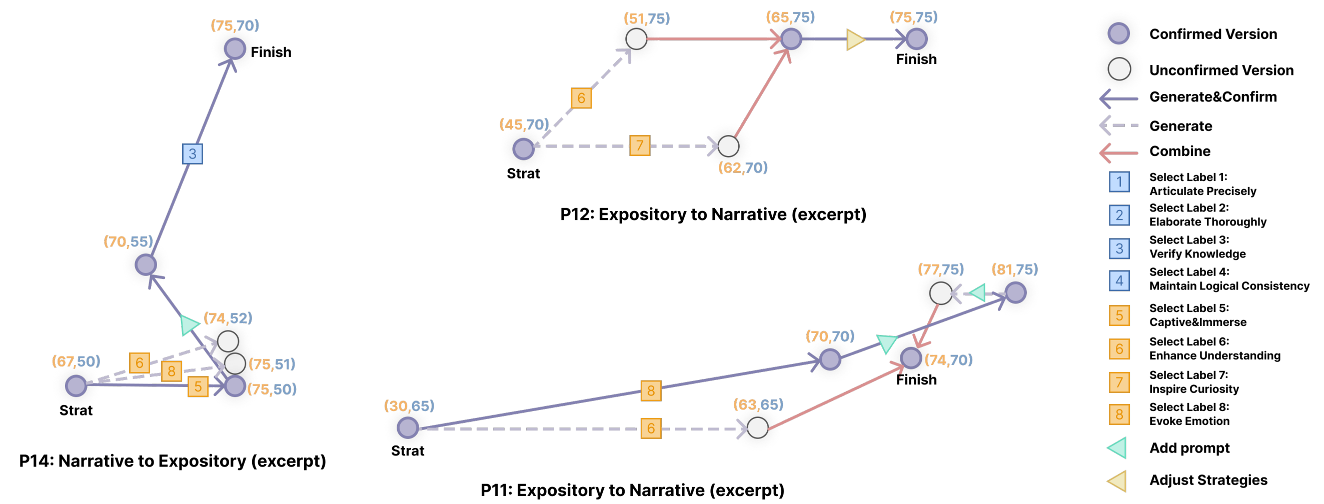

Figure 13: Visualization examples of segment revisions from P11, P12, and P14.

7.1 2D Coordinate Visualization for Balancing Scientific Accuracy and Narrative Engagement

7.1.1 Visualize the Communicative Goals The coordinate graph serves as a persistent, actionable reference that maps abstract writing goals and tradeoffs into a tangible representation. Each node on the graph represents a version selected by participants for evaluation based on two key communicative goals: scientific accuracy and narrative engagement. Most participants reported that the visualization facilitated their ability to prioritize their revisions. As P3 put it, "The coordinate graph is a feature that typical AI tools lack. It keeps me from getting lost of balancing the two dimensions during revisions." Furthermore, participants used the scores to prioritize their focus. As P12 said, "I refer to the scores to decide which dimension I need to improve." Similarly, P6 noted, "If the two dimensions differ too much, it reminds me to pay more attention to the other." By externalizing implicit internal writing goals, participants were able to engage in both self-monitoring and high-level planning. The system tends to facilitate metacognitive regulation by allowing users to visualize how each revision aligned with their tradeoffs, compare iterations, and identify areas for improvement.

Participants reported using the graph's visualization to make informed decisions about their revisions. As P8 shared, "I can see strengths and weaknesses by comparing the new node with the old one; if scientific accuracy drops, I adjust accordingly in the next generation." P10 added, "If I want the text to be more narrative, I just check if the engagement score of the newly generated node is higher than the previous one before reading the content carefully. With the baseline, I had to judge that on my own, and there was no version comparison to help me see which one was better." P16 also appreciated the visualization's clarity:

> *"When multiple nodes are generated by clicking different labels, I can intuitively compare them by observing their positions on the coordinate axes to see their scores in scientific accuracy and narrative engagement. This makes it easier to interpret the differences between nodes in a clear and direct way.".*

Such visual comparisons also helped reinforce participants' confidence in their editorial decisions. As P3 explained,

> *"The coordinate scores serve as a valuable reference point, helping me align my edits with my internal standards and feel more confident in the revisions I make, as I can visually see I am on my way to my desired direction. For example, when I aim to improve engagement, and I see that the engagement score increases, that reinforces my decision.".*

7.1.2 Visualization Drives Iteration The process of using the coordinate axes to assess current versions along the two dimensions constructively drove further iterations. As illustrated in **Appendix A.6(Figure 15)**, when attempting to add storytelling and narrative elements to expository content, participants initially selected labels associated with narrative engagement. However, during later iterations, they often returned to labels targeting scientific accuracy in order to restore balance.

This kind of iteration can also be observed in **Figure 13**. For example, in the case of P14, when she attempted to revise a text from a narrative storytelling version to one with more scientific expression and explanatory content, she initially selected the label Captivate & Immerse, along with other engagement-enhancing labels. After fine-tuning the text at that stage using prompts, she realized the need to further improve scientific accuracy. As a result, she selected the Verify Knowledge label and eventually accepted the final version. This shows how the coordinate axes helped her take both dimensions into account and negotiate a balance between them.

7.2 Supporting Metacognition in LLM Writing

7.2.1 Facilitating Metacognitive Knowledge and Control through Label-Based Scaffolding The use of structured labels supports both metacognitive knowledge (explicit awareness of strategies) and control (breaking down goals into manageable steps). When faced with open-ended writing tasks, users often struggle to identify effective strategies. Without clear guidance, the process of planning and task decomposition can feel overwhelming, especially for those with less writing experience. By providing clear guidance and reducing the effort of remembering or retrieving strategy knowledge, labels enable users to develop a deeper understanding of their own thought processes and take more effective actions.

As P11 described, “These eight labels give me directions; otherwise I wouldn’t know how to begin,” illustrating how labels helped transform an ambiguous task into a navigable one. Similarly, P5 viewed them as “hints” that sparked new ideas for enhancing text engagement. These labeled entry points externalized editorial heuristics, allowing users to shift from general intentions to concrete strategies. In addition, the system also reduced the effort to remember or retrieve the knowledge of the strategies. As P7 remarked, “With the baseline, I’d be too lazy to dig through an Excel sheet for strategies. Here, they’re just packaged.” P16 added, “I don’t need to remember what each function does. I just click and go—these labels offer a clear framework.”

Beyond easing metacognitive demands related to strategy choice and task decomposition, the labels also encouraged users to break habitual patterns and reflect on alternate approaches. “It gave me methods I hadn’t considered,” said P12. “I used to edit habitually, but this nudged me toward new directions.” This indicates how structured cues also served as catalysts for metacognitive control—users were not only executing known strategies, but also experimenting with new ones.

7.2.2 Improving Metacognitive Monitoring through "Muse" Reflective Feedback By leveraging feedback and prompting metacognitive monitoring, Muse helps its users develop a deeper understanding of their own thought processes and apply that knowledge to improve their writing. Effective metacognitive monitoring involves not just executing strategies, but noticing what is missing, what is working, and how one’s understanding is evolving. Muse provides users with reflective feedback that functions as a metacognitive mirror, helping them recognize overlooked patterns and strengths in their writing.

P1 recalled such a moment while revising an explanation of Archimedes’ principle: “A metaphor suggested by Muse struck me: the idea that buoyant force is equal to the displaced water’s weight, much like the balanced arms of a scale. This visual analogy illuminated the concept for me.” Such moments support both evaluation (assessing one’s output) and awareness (recognizing conceptual gaps), which are core aspects of metacognitive monitoring. This feedback also prompted the internalization of new editorial strategies. Similarly, P15 noted, “I started using strategies I hadn’t tried before, and I remembered to use them again.” These quotes suggest that system feedback helped convert momentary insights into lasting metacognitive knowledge.

Several participants described moments where the feedback didn’t just help with one revision but reframed how they thought about writing more broadly. As P6 said, “I started seeing where I tend to do well or poorly. Muse pointed out strengths I didn’t even realize I had.” This aligns with prior findings on self-monitoring: users must develop an awareness of their own tendencies before adjusting them. As P10 explained, “With more guidance during revision, I felt like I was internalizing a way of thinking. Even without the system, I’d know how to approach future writing.”

This finding aligns with the quantitative results that *Equinox* supports more on "reflecting on my writing strategies and choices " (M = 5.5 (*Equinox*) vs. 4.63 (Baseline), p = .013) (**Figure 11 Q3**) compared with the baseline.

7.2.3 Fostering Metacognitive Flexibility: Supporting Parallel Comparison and Exploration The system provides participants with a high degree of flexibility in revision, enabling them to explore multiple directions in balancing narrative engagement and scientific accuracy, while also supporting fine-tuned adjustments within a chosen axis. In contrast to the baseline system, which enforces a linear revision process, this canvas-based interface facilitates parallel comparison and ongoing exploration.

Participants described their interaction with the system as playful and exploratory. As P11 reflected, “I wanted to see how different strategies under the same label changed the output, so I generated multiple versions. It gave me room to play and test.” The system minimized cognitive and temporal overhead, allowing for a low-stakes, high-feedback interaction that encouraged curiosity and iteration.

Participants highlighted the value of viewing multiple alternatives simultaneously. As P6 noted, “These labels give me several options with different focuses at the same time in the canvas. I can choose one version to develop further and still return to earlier iterations after generating new branches.” This non-linear workflow allowed for reflective comparison and discouraged premature commitment to a single version.

Moreover, the system occasionally served as a catalyst for unexpected creativity. In one case, P11 recalled selecting the label “enhance understanding,” which led to the automatic insertion of a metaphor: “That metaphor was so on-point, I hadn’t even thought about that kind of revision before.” Such moments illustrate the system’s potential to support conceptual innovation, introducing rhetorical strategies beyond users’ initial expectations.

These findings aligns with the quantitative data that participants rated *Equinox* offer more flexibility to "adjust my writing strategies during the editing process" (M = 5.69 (*Equinox*) vs. 4.56 (Baseline), p = .0016)(**Figure 11 Q4**) and give more supportive of exploration support to generate "diverse ideas and outcomes " (M = 5.13 (*Equinox*) vs. 3.69 (Baseline), p = .004) (**Figure 10 Exploration**) compared to the baseline.

7.3 User Expectations for the Future of *Equinox*

7.3.1 The Tension Between 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, "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."

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." 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."

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. The interpretability of the scores also needs to be improved, as indicated by some participants.

7.3.2 Experienced Writers Seek More Flexible and Customizable Labels While the fixed label set was seen as a helpful starting point, some experienced users felt it could be expanded to better support their advanced needs. P3, a seasoned science communicator, shared: "The eight labels are a solid foundation, but I would appreciate a broader set to support more diverse explorations." P1, P3, P2, and P14, all of whom are experienced science communicators or experienced writers, expressed interest in more customizable labels, such as they can combining or tailoring underlying strategies to form customized labels to align more closely with their specific goals. 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." This indicates a demand for labels that can be tailored to individual needs.

7.3.3 Muse as a Future Co-Editor While participants appreciated what Muse could already do, many imagined what it might become. P2 wanted more real-time dialogue: "I wish it were more interactive—like chatting with someone who helps me reflect as I go." P14 hoped for more adaptability: "The more I use it, the more I want it to understand how I write and suggest things based on that." Others wished for more precision in the feedback. "Right now, Muse gives high-level suggestions," one participant said. "But it'd be more useful if it could point to which step or decision was strong or weak, and explain why." These comments suggest that participants saw Muse not just as a tool for generating or revising text, but as a partner that could grow with them—learning their writing style, giving relevant feedback, and helping them refine how they think through revisions.

8 Discussion

8.1 How Dual Mechanisms of Support Promote Metacognition in LLM-Assisted Writing

Following the dual-path framework proposed by Tankelevitch et al. [86], this section examines how *Equinox* supports metacognitive engagement in LLM-assisted writing through two complementary mechanisms. The first path focuses on improving users' metacognitive abilities by scaffolding core reflective processes such as planning, monitoring, and strategic control. The second path addresses the need to reduce metacognitive demand, by redistributing evaluative effort to the interface through visual cues and interaction design. These pathways—strengthening ability and relieving burden—work together to enable more deliberate, confident, and cognitively sustainable writing practices.

8.1.1 Fostering Metacognitive Capacities through Visual Feedback and Strategy Scaffolding Rather than merely assisting with text generation, the dual-axis coordinate design of *Equinox* helps users developmetacognitive capacities—including self-awareness, task decomposition, and evaluative control—through structured scaffolding. This corresponds to what Tankelevitch et al. [86] define as "metacognitive support strategies," which aim to improve the user's own ability to plan and manage their thinking.

Equinox's strategy labels further scaffold rhetorical decision-making by helping users break down abstract goals into concrete, achievable editing directions. This decomposition avoids the ambiguity often encountered when users must convey their intentions solely through free-form prompts, which can be misinterpreted or too vague for targeted revision [97]. By prompting users to select high-level revision intents—such as enhancing understanding or increasing credibility—the system structures revision into discrete, traceable moves that reflect intentional rhetorical planning. This not only reduces the cognitive load of spontaneous strategy formulation, but also supports users in articulating and pursuing their communicative objectives with greater clarity.

Beyond individual iterations, the revision tree invites reflective comparison and experimentation across versions. Users are encouraged to explore, return, and recombine edits, supporting metacognitive flexibility [82]. This not only reinforces adaptive control and pattern recognition across revisions, but also nurtures creativity by inviting contrastive reasoning—seeing how different strategies shape different rhetorical effects. In sum, *Equinox* cultivates an environment where users engage not only in writing, but in learning to manage their writing process more consciously.

8.1.2 Reducing Metacognitive Demands through Innovative Interface While scaffolding fosters user's metacognitive capacities,*Equinox*also reduces the cognitive cost of metacognition by embedding reflective structures directly into the interface. In line with Tankelevitch et al.'s [86]second pathway—reducing metacognitive demand through innovative interface design—this approach transforms high-effort reflection into low-friction visual interpretation.

The coordinate axis enables ambient feedback about the quality of outputs in relation to abstract tradeoffs during the revision process. Instead of evaluating outputs through close reading, users can rely on spatial cues to assess performance and decide next steps. This supports epistemic efficiency—using external structures to reduce internal computation—and enables rapid goal reorientation when attention or working memory is limited [85]. The revision lattice complements this by affording parallel comparison. Rather than tracking edits linearly, users can scan, contrast, and prioritize among alternatives without serial judgment. This fosters more confident decision-making and reduces the perceived ambiguity of LLM outputs—especially when exploring unfamiliar revision directions. Moreover, users employed heuristic zones on the coordinate graph to guide confidence and stopping decisions. This suggests that offloading can support not just immediate assessment, but also self-regulatory boundaries, allowing users to recognize when a revision is "good enough." Importantly, however, this redistribution of cognitive labor shifts the nature of reflection: from deliberate strategy execution to more intuitive, score-guided navigation.

This shift is not without risks. When system feedback becomes too legible or directive, users may defer too readily to system judgment, narrowing their engagement with content or limiting deeper evaluation. Indicated by our results, some uncertainty also remains around how scores are generated and how they relate to actual textual quality, revealing a persistent tension between efficiency and interpretability. While many users found the scaffolding helpful for clarifying revision direction—particularly when uncertain—there were also calls for greater customizability. Experienced creators, in particular, sought more flexible and composable strategy labels to better express their rhetorical intent. This resonates with prior frameworks that highlight the dual-edged nature of customizability [86]: it can empower user control, but also complicate guidance if not carefully managed, especially for novice users. It's not a simple decision and should be tailored and differentiated based on the user's confidence and expertise [14].

Future iterations may strengthen this balance by enabling tighter coupling between strategy labels and coordinate feedback—for example, making visible how specific rhetorical moves impact different evaluation dimensions. Enhancing the interpretability of scores and making underlying evaluation logic transparent could help users better calibrate trust and refine their editing decisions. Furthermore, expanding the label system to include customizable modules for broader writing strategies, such as grammar correction, tone adjustment, or evidence elaboration, would support more diverse workflows while preserving the benefits of strategic scaffolding. In this way, *Equinox* points toward a future where LLM writing tools not only support cognition, but evolve in tandem with it, enabling co-authorship that is both adaptive and deeply intentional.

8.2 How Spatial Visualization Facilitates Strategic and Exploratory Use of LLMs in Writing

8.2.1 From Sequential Drafting to Spatial Thinking: Rethinking LLM Writing Interactions As LLM writing tools become increasingly embedded in writing workflows, a persistent limitation is their reliance on linear, single-threaded interaction [81, 84]. Most systems guide users through sequential drafts, offering limited support for exploring multiple directions in parallel or revisiting earlier ideas with strategic intent. Even when some tools enable parallel exploration by generating multiple versions simultaneously, they typically lack a structured scaffolding framework to facilitate meaningful co-creation with the LLM. Moreover, they do not provide visual, real-time feedback on how each version progresses toward distinct writing goals.

Our design addresses this gap by introducing a coordinate-based visualization system that maps revisions along interpretable axes to balance scientific accuracy and narrative engagement. Writers are not just reacting to LLM output; they are shaping a landscape of rhetorical possibilities. The system enables easy generation, comparison, and synthesis of divergent drafts, supporting a more deliberate and strategic writing process. In user studies, participants described using the visualization to spread out their thinking and retrieve prior edits when exploring new rhetorical strategies. Crucially, this exploratory functionality is grounded in cognitive and learning sciences.

Modular and extensible, the design holds strong potential as a core component in the toolbox of LLM-based writing products. Rather than prescribing fixed goals or workflows, it can invite users to define their own axes of evaluation—enabling visualization of choices and trade-offs across diverse writing scenarios. This flexibility empowers users to engage in parallel thinking and make more intentional, informed decisions with LLM outputs to meet their writing goals, whether in academic, creative, or collaborative contexts. As part of a larger ecosystem, it lays the groundwork for a new generation of writing tools centered on strategic, user-directed co-creation.

8.2.2 Toward a Generative Canvas: Expanding the Design Space of Coordinate-Based Writing The underlying coordinate-based interaction paradigm holds broader potential for domains that demand more flexible, exploratory authoring creative support. In contexts such as fiction, poetry, or screenwriting, where goals like emotional resonance, narrative pacing, or stylistic novelty are prioritized, the coordinate space can serve as a dynamic interface for decision-making when co-creation with LLM. We outline six key design considerations for extending this paradigm. Together, these themes reimagine writing not as a linear pipeline but as a spatial, interactive canvas for thought and transformation. As authoring tools evolve, such paradigms may offer more expressive and cognitively aligned experiences for human-AI collaboration. **Figure 14** visualizes the design space of Coordinate-Based Writing.

Score-Aligned Drag-to-Generate Revisions (A) Users can revise their text by simply dragging a version's node to a different point on the coordinate graph—such as increasing clarity from 60 to 90. The system then generates a new version that reflects this target score. This interaction makes the link between evaluation and revision more intuitive, letting users improve text quality just by moving a point, instead of writing a new prompt.

Hierarchical and Multi-Resolution Coordinate Spaces (B) Beyond flat representations, coordinate systems can be expanded into hierarchical structures. A user could zoom into a node to reveal a sub-coordinate space, enabling exploration of micro-revisions (e.g., different phrasings or sentence structures) within broader narrative shifts. This layered interaction supports both high-level planning and low-level editing in a unified spatial framework.

Flexible Axis Definition and Reconfiguration Multidimensional Coordinate Axes (C) Users can dynamically define or switch between different coordinate axes, adapting the interaction to evolving writing needs. For example, in addition to narrative tone and clarity, axes could reflect orthogonal dimensions such as grammar correctness, sentence length, or reader engagement. Multi-dimensional and even 3D coordinate spaces can accommodate diverse writing goals without being bound to genre-specific templates.

Direct Manipulation and Multi-Selection Editing (D) Inspired by design tools like Photoshop, users could directly manipulate one or more nodes together, selecting, dragging, or aligning them in bulk. Adjustments via sliders or handles could apply to

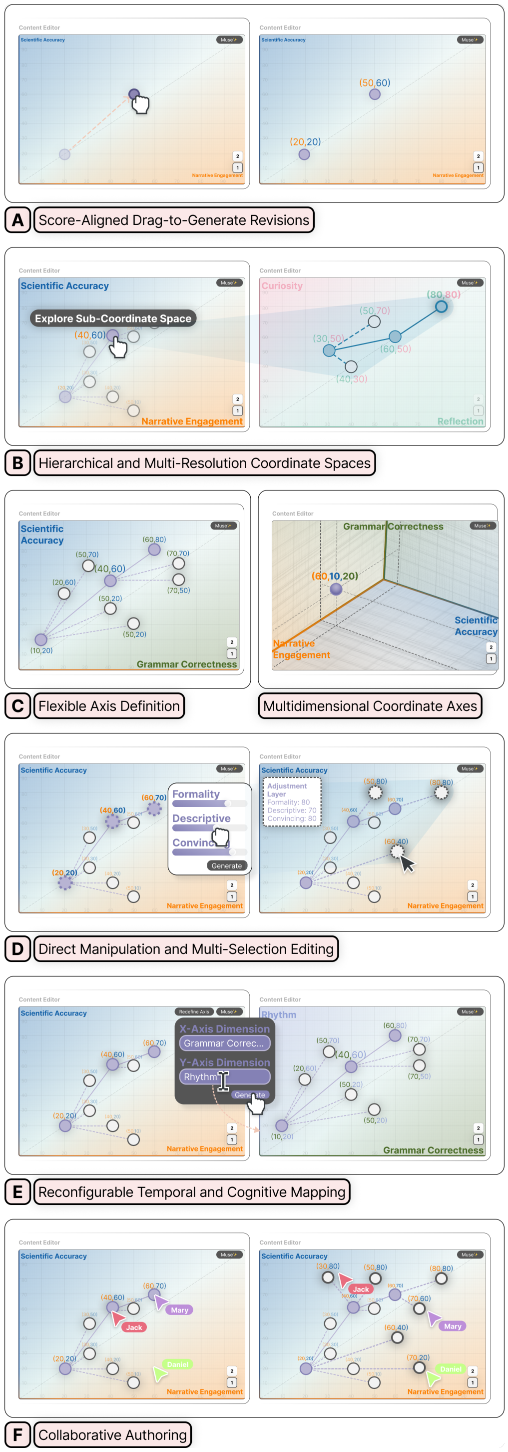

Figure 14: Visualization of coordinate-based interaction paradigms that re-imagine writing as a spatial and temporal iteration process

a single node or across a group, enabling global transformations (e.g., increasing all selected nodes' formality) or fine-tuned batch editing.

Reconfigurable Temporal and Cognitive Mapping (E) Axes can be redefined across time as the writer's focus shifts. For example, after completing structural edits, a user may redefine axes to evaluate rhythm and grammar. Older nodes can be remapped within new axis contexts, making the coordinate space a living archive of intent and cognition over time.

Collaborative Authoring in Shared Spatial Contexts (F) The coordinate interface enables structured collaboration by allowing multiple contributors to work in parallel across different regions of the space. Each contributor's inputs can be visualized dimensionally, facilitating merge, comparison, and synthesis. The coordinate tool can also integrate as a plugin into existing writing platforms, enabling co-creation without workflow disruption.

8.3 Limitation and Future Work

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

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

Evaluation Dependency on Proxy Scores Although *Equinox* provides real-time feedback on scientific accuracy and narrative engagement, this feedback is generated by a model trained on proxy metrics (e.g., perceived credibility and engagement from non-experts). While useful, these proxies may not fully capture the nuance of effectiveness in real-world science communication. Actual audience reactions in diverse contexts (e.g., classroom learning vs. YouTube videos) may differ from model predictions. Therefore, the reliability and generalizability of the scoring system should be validated further.

Additionally, 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.

9 Conclusion

We presented *Equinox*, a writing interface that foregrounds visual exploration and iterative refinement to support science communication. Through a dual-axis visualization and strategy-guided revision workflow, the system helps users navigate the trade-off between scientific accuracy and narrative engagement. Our study shows that this visual and iterative approach enhances metacognition and encourages creative exploration. *Equinox* demonstrates how real-time visualization can support iterative revision with LLM toward communicative goals.

References

[1] Badri Adhikari. 2023. Thinking beyond chatbots' threat to education: Visualizations to elucidate the writing or coding process. *Education Sciences* 13, 9 (2023), 922.

[2] TM Al-Jarrah, JM Al-Jarrah, RH Talafhah, and I Bashir. 2019. Exploring the effect of metacognitive strategies on writing performance. *Global Journal of Foreign Language Teaching* 9, 1 (2019), 33–50.

[3] J Craig Andrews and Terence A Shimp. 2018. *Advertising, promotion, and other aspects of integrated marketing communications*. Cengage Learning.

[4] Isabelle Augenstein. 2021. Determining the credibility of science communication. *arXiv preprint arXiv:2105.14473* (2021).

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

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

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

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

[9] Karen Bultitude. 2011. Science communication–Why and how? (2011).

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

[11] Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology* 12, 4 (2009), 321–347.

[12] 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.

[13] Stuart K Card, Jock D Mackinlay, and George G Robertson. 1991. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 99–122.

[14] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2022. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092* (2022).

[15] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.

[16] 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.

[17] 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*. 1–19.

[18] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Toyteller: AI-powered Visual Storytelling Through Toy-Playing with Character Symbols. *arXiv preprint arXiv:2501.13284* (2025).

[19] 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.

[20] 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. https://doi.org/10.1073/pnas.1320645111 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1320645111

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

[22] Dangin Dangin. 2020. Students' Awareness of Metacognitive Reading Strategies in Academic Reading. *Journal of English Teaching and Learning Issues* 3, 1 (2020), 33–42.

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

[24] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[25] Ana Delicado, Jussara Rowland, and Joao Estevens. 2021. Bringing back the debate on mediated and unmediated science communication: the public's perspective. *Journal of Science Communication* 20, 3 (2021), A10.

[26] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems* 37 (2024), 19783–19812.

[27] Julie S. Downs. 2014. Prescriptive scientific narratives for communicating usable science. *Proceedings of the National Academy of Sciences* 111, supplement_4 (Sept. 2014), 13627–13633. https://doi.org/10.1073/pnas.1317502111 Publisher: Proceedings of the National Academy of Sciences.

[28] Grant Eckstein, Jessica Chariton, and Robb Mark McCollum. 2011. Multi-draft composing: An iterative model for academic argument writing. *Journal of English for academic purposes* 10, 3 (2011), 162–172.

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

[30] Wiebke Finkler and Bienvenido Leon. 2019. The power of storytelling and video: a visual rhetoric for science communication. *Journal of Science Communication* 18, 05 (Oct. 2019), A02. https://doi.org/10.22323/2.18050202

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

[32] John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10 (1979), 906.

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

[34] Kexue Fu, Ruishan Wu, Yuying Tang, Yixin Chen, Bowen Liu, and RAY LC. 2024. " Being Eroded, Piece by Piece": Enhancing Engagement and Storytelling in Cultural Heritage Dissemination by Exhibiting GenAI Co-Creation Artifacts. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2833–2850.

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

[36] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. 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, 1–12. https://doi.org/10.1145/3290605.3300526

[37] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[38] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009).

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

[40] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* 3 (2023).

[41] Sandra G Hart. 1986. NASA task load index (TLX). (1986).

[42] Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory* 3 (2007), 265–289.

[43] 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.

[44] 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 (Oct. 2020). https://doi.org/10.3389/fcomm.2020.581349 Publisher: Frontiers.

[45] 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. https://doi.org/10.3233/SHTI220546

[46] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–20.

[47] 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.

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

[49] 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. https://doi.org/10.1525/collabra.18898 arXiv:https://online.ucpress.edu/collabra/article-pdf/7/1/18898/835600/collabra_2021_7_1_18898.pdf

[50] Anjali Khurana, Hariharan Subramonyam, and Parmit K Chilana. 2024. Why and when llm-based assistants can go wrong: Investigating the effectiveness of prompt-based interactions for software help-seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 288–303.

[51] 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. https://doi.org/10.1145/3563657.3595996

[52] 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*. 115–135.

[53] Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence* 6 (2024), 100225.

[54] 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

[55] 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. https://doi.org/10.1093/biosci/biu084 arXiv:https://academic.oup.com/bioscience/article-pdf/64/8/719/8719054/biu084.pdf

[56] Joe Lambert. 2013. *Digital storytelling: Capturing lives, creating community.* Routledge.

[57] Robert A Lehrman and Eric Schnure. 2019. *The Political Speechwriter's Companion: A Guide for Writers and Speakers.* CQ Press.

[58] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760* (2023).

[59] Norma J Livo and Sandra A Rietz. 1986. Storytelling: Process and practice. *(No Title)* (1986).

[60] 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. https://doi.org/10.1145/3643834.3661587

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

[62] Damien Masson, Young-Ho Kim, and Fanny Chevalier. 2024. Textoshop: Interactions Inspired by Drawing Software to Facilitate Text Editing. *arXiv preprint arXiv:2409.17088* (2024).

[63] Daniel Gary McDonald Daniel Gary McDonald. 2014. Narrative research in communication: key principles and issues. *Review of Communication Research* 2 (2014), 115–132.

[64] Julia Metag, Florian Wintterlin, and Kira Klinger. 2023. science communication in the digital age—new actors, environments, and practices. *Media and Communication* 11, 1 (2023), 212–216.

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

[66] Thomas O Nelson. 1990. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*. Vol. 26. Elsevier, 125–173.

[67] National Academies of Sciences, Medicine, Division of Behavioral, Social Sciences, Committee on the Science of Science Communication, and A Research Agenda. 2017. Communicating science effectively: A research agenda. (2017).

[68] Reham Omar, Ishika Dhall, Panos Kalnis, and Essam Mansour. 2023. A universal question-answering platform for knowledge graphs. *Proceedings of the ACM on Management of Data* 1, 1 (2023), 1–25.

[69] Roger A Pielke Jr. 2007. *The honest broker: Making sense of science in policy and politics.* Cambridge University Press.

[70] Alastair Pollitt. 2012. Comparative judgement for assessment. *International Journal of Technology and Design Education* 22, 2 (2012), 157–170.

[71] 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.

[72] Fatemeh Rabiee. 2004. Focus-group interview and data analysis. *Proceedings of the nutrition society* 63, 4 (2004), 655–660.

[73] Marissa Radensky, Daniel S Weld, Joseph Chee Chang, Pao Siangliulue, and Jonathan Bragg. 2024. Let's Get to the Point: LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts. *arXiv preprint arXiv:2406.10370* (2024).

[74] 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

[75] 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.

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

[77] Margaret A Rubega, Kevin R Burgio, A Andrew M MacDonald, Anne Oeldorf-Hirsch, Robert S Capers, and Robert Wyss. 2021. Assessment by audiences shows little effect of science communication training. *Science Communication* 43, 2 (2021), 139–169.

[78] Keegan Sawyer and Brooke Smith. 2024. Communication and engagement for basic science: insights and practical considerations. *Journal of Science Communication* 23, 7 (2024), Y01.

[79] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary educational psychology* 19, 4 (1994), 460–475.

[80] Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207* (2024).

[81] Momin Siddiqui, Roy Pea, and Hari Subramonyam. 2025. Script&Shift: A Layered Interface Paradigm for Integrating Content Development and Rhetorical Strategy with LLM Writing Assistants. *arXiv preprint arXiv:2502.10638* (2025).

[82] Rand J Spiro, Paul J Feltovich, Michael J Jacobson, and Richard L Coulson. 2012. Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. In *Constructivism in education*. Routledge, 85–107.

[83] 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*. 1–19.

[84] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[85] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885* (2022).

[86] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

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

[88] Jason Toy, Josh MacAdam, and Phil Tabor. 2024. Metacognition is all you need? using introspection in generative agents to improve goal-directed behavior. *arXiv preprint arXiv:2401.10910* (2024).

[89] RY Tyaningsih, TW Triutami, D Novitasari, NP Wulandari, and YM Cholily. 2020. The relationship between habits of mind and metacognition in solving real analysis problems. In *Journal of Physics: Conference Series*, Vol. 1663. IOP Publishing, 012053.

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

[91] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. " It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–26.

[92] Qian Wan, Jiannan Li, Huanchen Wang, and Zhicong Lu. 2025. Polymind: Parallel Visual Diagramming with Large Language Models to Support Prewriting Through Microtasks. *arXiv preprint arXiv:2502.09577* (2025).

[93] 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*. 1–15.

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

[95] Dustin J Welbourne and Will J Grant. 2016. Science communication on YouTube: Factors that affect channel and video popularity. *Public understanding of science* 25, 6 (2016), 706–718.

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

[97] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[98] Haijun Xia, Hui Xin Ng, Zhutian Chen, 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*. ACM, New York City NY USA, 163–174. https://doi.org/10.1145/3491140.3528279

[99] Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. 2021. A design space for applying the freytag's pyramid structure to data stories. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 922–932.

[100] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[101] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

[102] Yu Zhang, Kexue Fu, and Zhicong Lu. 2025. RevTogether: Supporting Science Story Revision with Multiple AI Agents. *arXiv preprint arXiv:2503.01608* (2025).

[103] 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*. 1–18.

[104] 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*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581476

[105] 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*. 1–30.

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

A Appendix

A.1 Specific Strategies for Science Communication Writing

Table 3: Design Space for Science Communication Writing

Category	Strategy	Definition	Label
Scientific Accuracy	(1) Layered Transitions [54, 61, 76, 90]	Use multiple transition words or phrases (e.g., "but," "and," "therefore") within a short span to emphasize logical shifts and contrasts.	4
	(2) Rigorous Source Verification [4, 54, 74]	Cross-check scientific claims and data against reliable, peer-reviewed sources to ensure accuracy.	3
	(3) Step-by-Step Explanation [5, 54]	Introduce the core idea first and then progressively add background details, creating a structured learning process.	2, 4
	(4) Acknowledge Uncertainties [69]	Transparently discuss uncertainties, potential biases, or limitations in data and models to build credibility.	1, 2
	(5) Consistent Terminology [55]	Use the same terminology throughout the content to maintain clarity and avoid confusion.	1
	(6) Citations & Quotes [4, 29]	Integrate citations and direct quotes seamlessly to enhance credibility while maintaining narrative flow.	3
	(7) Everyday Events to Scientific Insights [5, 55]	Automatically identify and link theories or knowledge to real-world events or stories mentioned in the text.	2, 3
Narrative Engagement	(8) Question-Answer Hook [30, 44, 56]	Ask a direct question and provide an immediate answer to introduce key concepts clearly and concisely.	5, 6, 7
	(9) Reflection Question [30]	Ask a thought-provoking question that does not require an immediate answer, encouraging reflection and reinforcing key concepts.	5, 7, 8
	(10) Suspense-Driven Reveal [98, 104]	Present a question, problem, or scenario at the beginning and delay its resolution to sustain curiosity.	5, 7
	(11) Use metaphors [27, 30, 55]	Convey unfamiliar concepts by drawing analogies to more familiar ones.	5, 6
	(12) Inject humor [39]	Use playful language or puns to make the content more engaging and enjoyable.	5, 8
	(13) Add real-world supporting examples [57, 59]	Illustrate abstract concepts using relatable, real-world examples.	5, 6
	(14) Add stories [20, 21, 59]	Use narratives with characters, settings, and plot progression to enhance engagement and memorability.	5, 6, 8
	(15) Add an imagery description [3, 30, 38]	Use vivid, sensory details to help the audience visualize concepts.	5, 6
	(16) Create negative emphasis for focused attention [30, 38, 44, 65]	Highlight extreme negative outcomes to intensify focus and reinforce key lessons.	5, 8
	(17) Make positive emotion to expand action repertoire [30, 35, 38, 65, 75, 94]	Use uplifting messages, particularly in conclusions, to inspire optimism and motivation.	5, 8
Both	(18) Simplify and abstract language [45, 49, 106]	Rephrase complex scientific terminology or detailed descriptions into more general, accessible language without compromising core accuracy.	1, 6
	(19) Clarify Key Terms [65, 76]	Define complex or specialized terms at the beginning to establish a shared understanding.	1, 6
	(20) Key Point Recap [30, 65, 87]	Summarize the main points concisely at the conclusion of the content to reinforce memory retention.	1, 4, 6
	(21) Repeat key point(s) or question(s) [6, 48]	Reinforce key concepts by strategically repeating crucial terms or questions.	1, 6
	(22) Emphasize with Numbers [33, 99]	Connect scientific discussions to real-world recent news or trends to enhance relevance and engagement.	1, 2, 3, 8
	(23) Strengthen the Connections Between Content [61, 90]	Ensure smooth transitions between related ideas by using bridging statements or contextual links.	4, 6
	(24) Present Balanced Views [55]	Provide both supporting evidence and counterarguments to present a well-rounded discussion.	2, 6
	(25) Tie Science to Current Events [5, 55]	Connect scientific discussions to real-world recent news or relavant stories.	3, 5, 6

Lable:Scientific Accuracy Effects:* 1. Articulate Precisely; 2. Elaborate Thoroughly; 3. Verify Knowledge; 4. Maintain Logical Consistency
Narrative Engagement Effects: 5. Captivate & Immerse; 6. Enhance Understanding; 7. Inspire Curiosity; 8. Evoke Emotion

A.2 Rating Model Construction

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

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

Score Collection We designed a survey to collect ratings for these texts on two dimensions: Narrative Engagement and Scientific Accuracy, two main communication goals in popular science [19]. For Narrative Engagement, we used five subscales: Narrative Presence, Emotional Engagement, Narrative Understanding, Curiosity, and General Narrative Engagement, a survey developed by prior work [11]. For Scientific Accuracy, given the lack of mature scales, we measured five dimensions inspired by standards for scientific texts from previous research [19]: Conceptual Clarity, Plausibility, Completeness, and Factual Correctness. When it comes to scientific accuracy, our focus is more on the audience's subjective experience during reading rather than an objective verification of accuracy. Since readers vary in their background knowledge, what we emphasize is not just factual correctness, but the perceived trustworthiness of how the content is presented — that is, how reliable and credible the text appears to them The full questionnaire can be found in the section .

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

Survey Results The distribution of scores for the 45 texts is displayed in the **Figure 9**. It is shown that story-like texts tend to elicit higher narrative engagement but exhibit lower scientific accuracy. In contrast, expository texts maintain higher scientific accuracy at the expense of engagement. The infotainment style appears to strike a balance between the two. Additionally, longer texts generally perform better in both dimensions, whereas shorter texts show lower overall scores, likely due to limitations in content depth and development.

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

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

A.2.1 *Example of Content*

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

A.2.2 *Survey used for gathering audience feedback*

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

A.2.3 *Score Criteria*

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

A.3 Materials used for experiment

Please view the materials via this anonymous link: https://cryptpad.fr/doc/#/2/doc/view/Q3Jhj+HhzHtt9zYqyF0Sv4mziQYBp6oWl43a84Gqmeg/

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

A.4 Survey

Part 1: Metacognition

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

Q1: I am aware of my writing goals during the editing process.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

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

Q2: I set specific goals for what I wanted the narrative to achieve.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q3: I reflect on my writing strategies or editing choices while using the AI writing tool. (Indicates real-time assessment of strategy effectiveness.)
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q4: During writing, I regularly checked whether the narrative was staying on track with my intended message.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q5: I can clearly identify areas of my writing that need improvement when using the AI tool.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q6: After writing, I reviewed the narrative to assess how well it communicated the scientific content.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q7: I am able to adjust my writing strategies during the editing process.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 2: Control (Control:)

Q8: I felt in control of the writing process while interacting with the system.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q9: I was able to override or ignore the system's suggestions when I thought it was necessary.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q10: I determined the direction and flow of the science narrative, not the system.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Part 3: Autonomy (Autonomy:)

Q11: I felt free to make my own choices during the co-writing process with the system.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q12: The system supported my ability to express my own ideas in the narrative.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Q13: I did not feel pressured to accept the system's suggestions.
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

A.5 Participants demographic information

.

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

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

A.6 User Interation data

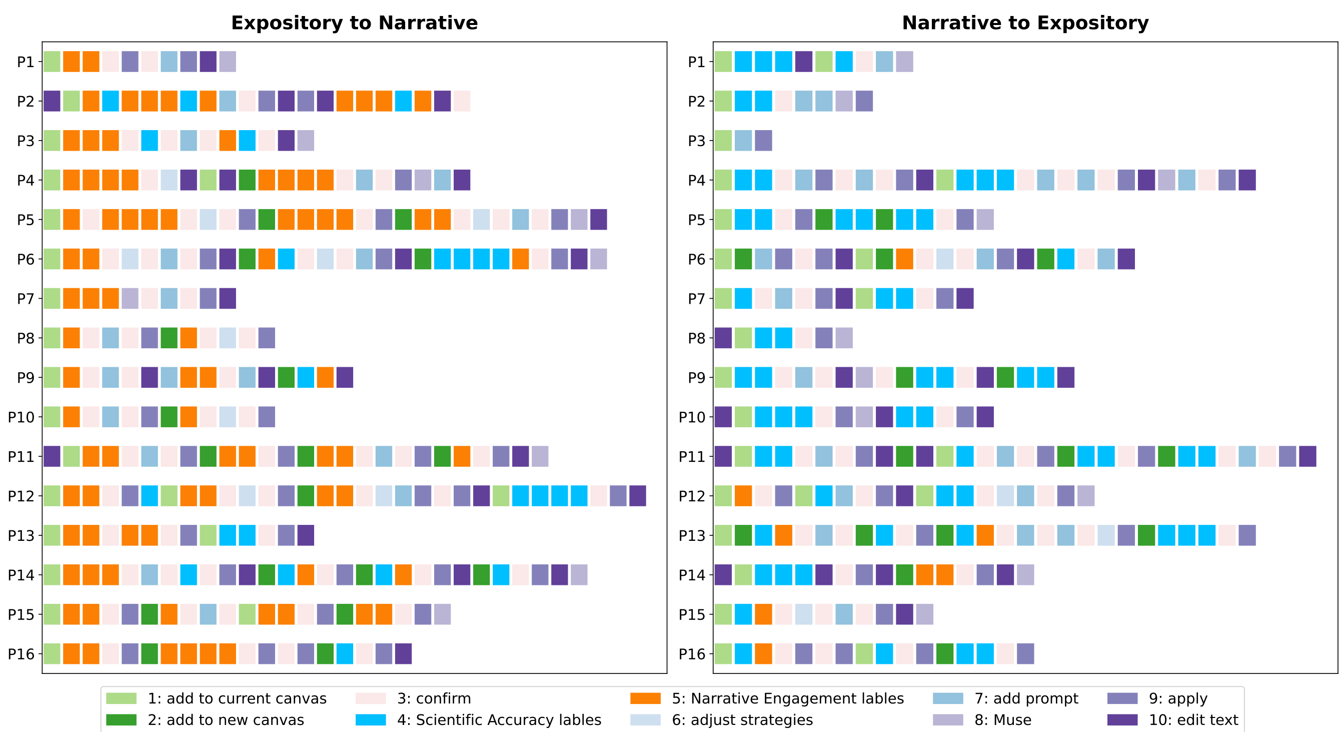

Figure 15: Visualization of interaction behaviors from 16 participants across two revision directions.

A.7 Prompts

A.7.1 Recommender

The blue word will be replaced by input information.

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

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

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

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

A.7.2 Generator

The blue word will be replaced by input information.

Generate new text based on user selected goals

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

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

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

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

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

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

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

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

Regenerate new text based on user feedback

```
# Order prompt 1
You are an expert in science communication and text refinement.
Your task is to modify the given text based on the provided instructions.
Step 1: Analyze the Input- Review the current text: {text}.

# Order prompt 2

Step 2: Remove the strategies, and the definitions is {strategy_info}.
Please modify the text by canceling the updated strategies while maintaining clarity and coherence.

# Order prompt 3
And please adjust the text according to the given instructions {user_prompt} from user while preserving its original meaning.
Adjust and modify the generated content completely based on the feedback from users

# Order prompt 4
Ensure that modifications enhance engagement, readability, and scientific accuracy.
- Maintain logical flow and avoid excessive lengthening or shortening of the text.

Step 3: Output the Updated Version Return the improved text in JSON format as follows:
{
"summary": "How the new generated version is different from the previous version",
"newText": "This is the refined version of the text, updated based on the strategy changes and/or custom prompt."
}
Make sure the newText is different from the original text.
Ensure that the JSON is properly formatted and contains no extra text.
```

Combine multi modified texts

```
# Order prompt
Step 1: Analyze the Input Texts
- Examine the content of texts {combine_test_list}.
- Identify the core scientific message of both.
- Extract the strategies used for each text from their strategy: {combine_trategy_list}.

Step 2: Integrate Strategies and Content
- If both texts use similar strategies, reinforce those elements for greater impact.
- If different strategies are used, merge them effectively.
- If text1 focuses on clarity(simplifying complex terms) and text2 focuses on engagement (adding analogies).
  integrate both elements.
- Ensure logical flow, avoiding abrupt shifts in tone or complexity.
- Keep the overall length nearly unchanged while ensuring coherence.

Step 3: Output the Combined Version Return the final output in the following JSON format:
{
"summary": "Summarization of how these different versions have combined together",
"newText": "This is the final merged version of text1, text2 and text X"
}
Ensure that the JSON is properly formatted and contains no extra text.
```

A.7.3 Scorer

The blue word will be replaced by input information.

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

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


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

# Order prompt 1
This is the original text: {text} and its score {currentScore}. Please use this as a reference.
Compare the current version with the original one in terms of scientific accuracy and narrative engagement, and assess
whether it performs better or worse than the previous version.
Compared to the previous version's scores, assign a score difference within a reasonable range.

# Order prompt 1
This is the {newText} you should evaluate. Return a JSON list in the format
{
"score": [The score of Narrative Engagement, The score of Scientific Accuracy]
}
Let's think step by step.
Please don't give a zero score in these two dimensions.

# Order prompt 3
This is the original texts: {combine_test_list} and their corresponding score list scoreList.
Please use this as a reference.
Compare the current version with the original one in terms of scientific accuracy and narrative engagement, and assess
whether it performs better or worse than the previous version.
Compared to the previous version's scores, assign a score difference within a reasonable range.
Please don't give a zero score in these two dimensions.
```

A.7.4 Analyzer prompt

The blue word will be replaced by input information.

Summary user chat history.

```
# Base prompt
The user is editing science communication narratives with the goal of balancing scientific accuracy and narrative
engagement.

Step 1: Understand Background (for context only).

Goals of making good science narratives:
Scientific Accuracy
- Factual correctness: Scientifically sound and valid - Clarity: Definitions and explanations are easy to understand
- Contextualization: Places information within appropriate scientific context
- Balanced perspectives: Includes both benefits and limitations
- Avoid oversimplification: Simplifies without distorting meaning.
Narrative Engagement
- Hooks: Starts with questions, vivid scenes, or curiosity-inducing facts
- Storytelling: Uses anecdotes, characters, or imagined scenarios
- Emotion: Inspires empathy, reflection, or awe
- Personal relevance: Encourages the reader to relate or reflect
- Flow: Smooth transitions and rhythm to aid readability
- Curiosity triggers: Surprising statistics or contrasts.
Trade-offs: More storytelling/emotion may reduce clarity or accuracy.
Heavier facts/context may feel dry or hard to follow.
Good narratives selectively balance both based on the intended audience and goals.


Step 2: Your Task You are given a record of how the user edited the narrative over time.
Each history node contains:
- score: A pair of numeric scores (e.g., [Narrative engagement, Scientific accuracy])
- isConfirmed: Whether the user accepted this version
- shortSummary: A brief note on the changes in how the strategy is been used
- fullText: The full revised version based on the label and strategy been used
- userPrompt: Instructions given by the user (optional) for further editing based on this node
- attitude: User stance (Normal, Like, Dislike) for the current node
- currentLabels: Conceptual goals/tags guiding this revision
- usedStrategies: Writing/editing strategies used.
- source: Source node ID Represents the parent node of the generated content.
- target: Target node ID, the current node
- type: Type of transformation from source node ID to the target node ID (e.g., Generate, Regenerate).

Now, based on this, complete the following three analyses:
1: Behavior Pattern & Preferences Analyze the user's editing patterns, preferences, and tendencies
(confirmation status, strategy usage, like or dislike this nodes, whether to perform the next operation under the current
node to generate child nodes etc.).
What are the user's habits and editing style? What are the strengths and weaknesses of their process? Are they achieving
a balance between engagement and accuracy?
2: Opportunities for New Directions and Strategies Review all versions (final and intermediate).
Identify unused or underused labels or strategies that might enrich the narrative.
Offer targeted improvement suggestions based on the user's specific edits, and summarize potential areas for further
refinement of the current text.

# Order prompt
This data is stored in: {chat_history}.
Return a structured summary with these sections:
1. User Preference & Behavior Pattern, including advantages and weakness.
Evaluate whether the current version meets the goals of science communication, and identify areas that could be improved.
2. New Strategy Suggestions for the content from our strategy list.
This is the list of strategies under each label: {all_strategy}.
Identify suitable strategies for improving the science communication narrative that were not used by the user in the chat
history, or were previously suggested but not effectively applied.
Justify each recommendation with clear reasoning.
Let's think step by step.
```

Update user preference based on user feedback.

```
# Base prompt
This is the user's history analysis along with their own feedback.
Please take into account the user's past behavior, preferences, and their analysis of your generated results
when recommending strategies:

# Order prompt
Based on the summarized user behavior analysis and strategic recommendations, including user preferences {user_preference}
and feedback {feedback}.
Return the result—integrating the user's feedback—for consideration by another AI agent to update
its strategy recommendations.
```

A.7.5 Filter

The blue word will be replaced by input information.

```
# Order prompt
You are an intelligent text filter.
I have provided several modified versions along with their scores. Your job is to select the versions with higher scores.
If two versions have similar scores, choose only the better one.
Retain exactly {number} texts.
Here is the {filter_prompt}.
Return a JSON list in the format
{
selected_result": [true, false, ...]
}
The list length must equal the total number of texts provided, with exactly number true values.
```