Synthia: Visually Interpreting and Synthesizing Feedback for Writing Revision

Chao Zhang
cz468@cornell.edu
Cornell University
Ithaca, NY, USA

Kexin Ju
kexinju@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Zhuolun Han
zh429@cornell.edu
Cornell University
Ithaca, NY, USA

Yu-Chun Grace Yen
yyen@cs.nycu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan

Jeffrey M. Rzeszotarski
jeffrz@cornell.edu
Cornell University
Ithaca, NY, USA

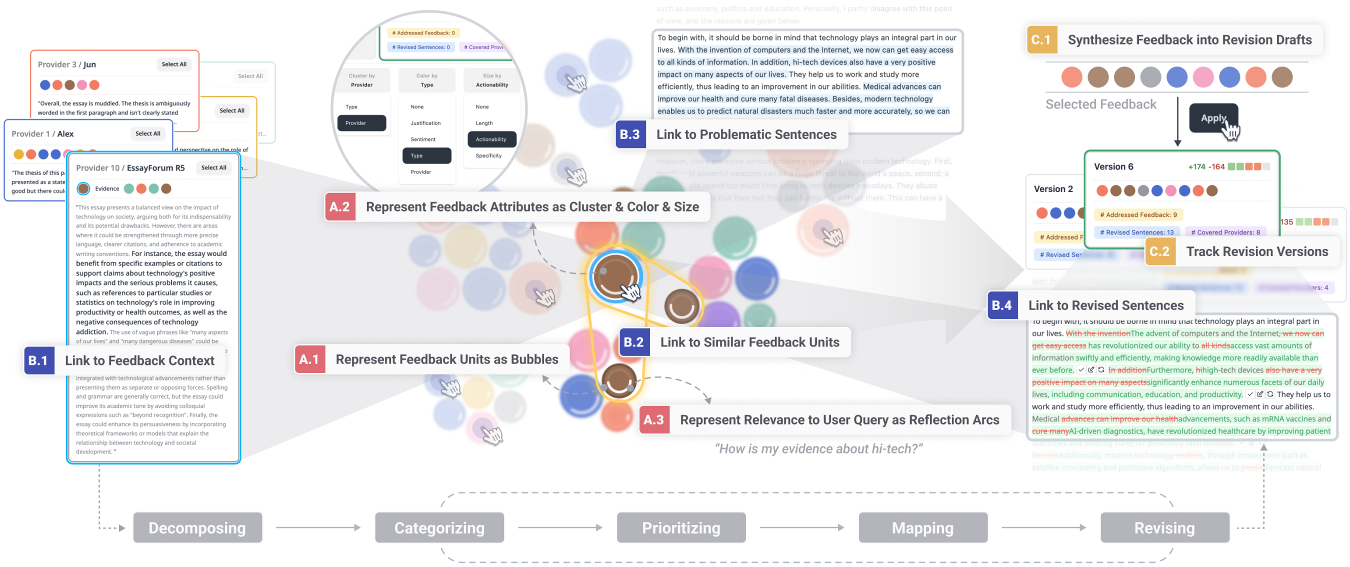

Provider 3 / Jun
Select All
Provider 1 / Alex
Provider 10 / EssayForum R5
Evidence
B.1 Link to Feedback Context
Cluster by Provider
Color by Type
Size by Actionability
A.2 Represent Feedback Attributes as Cluster & Color & Size
A.1 Represent Feedback Units as Bubbles
B.2 Link to Similar Feedback Units
A.3 Represent Relevance to User Query as Reflection Arcs
"How is my evidence about hi-tech?"
B.3 Link to Problematic Sentences
C.1 Synthesize Feedback into Revision Drafts
Selected Feedback
Apply
Version 2
Version 6
# Addressed Feedback: 9
# Revised Sentences: 13
# Covered Providers: 8
C.2 Track Revision Versions
B.4 Link to Revised Sentences
Decomposing
Categorizing
Prioritizing
Mapping
Revising

Figure 1: This figure illustrates how Synthia supports users in managing, interpreting, and synthesizing feedback for writing revisions. Synthia decomposes large collections of feedback into interactive, configurable bubbles (A.1), allowing users to adjust visual encodings, such as clustering, color, size, and reflection arcs, to identify patterns and assess feedback helpfulness (A.2 & A.3). Bidirectional highlighting connects feedback to its original context (B.1), similar feedback units (B.2), and the relevant sentences in the text (B.3), helping users understand its meaning, relevance, and impact. Finally, Synthia facilitates revision by enabling users to synthesize feedback into new drafts (C.1) and track revision versions (C.2), providing a non-linear, traceable interface for exploring alternative edits.

Abstract

While recent advances in HCI and generative AI have improved authors' access to feedback on their work, the abundance of critiques can overwhelm writers and obscure actionable insights. We introduce Synthia, a system that visually scaffolds feedback-based writing revision with LLM-powered synthesis. Synthia helps authors strategize their revisions by breaking down large feedback collections into interactive visual bubbles that can be clustered, colored, and resized to reveal patterns and highlight valuable suggestions. Bidirectional highlighting links each feedback unit to its original context and relevant parts of the text. Writers can selectively combine feedback units to generate alternative drafts, enabling rapid, parallel exploration of revision possibilities. These interactions support feedback curation, interpretation, and experimentation throughout the revision process. A within-subjects study ($N = 12$) showed that Synthia helped participants identify more helpful feedback, explore more diverse revisions, and revise with greater intentionality and transparency than a GPT-4-based writing interface.

UIST '25, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747703

CCS Concepts

• **Human-centered computing → Interactive systems and tools**; **Natural language interfaces**; **Information visualization**.

Keywords

Feedback, Sensemaking, Writing, Revision, Visual Interfaces, Human-AI Interaction

ACM Reference Format:
Chao Zhang, Kexin Ju, Zhuolun Han, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Synthia: Visually Interpreting and Synthesizing Feedback for Writing Revision. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3746059.3747703

1 Introduction

Revision is essential for high-quality writing. It often involves gathering and synthesizing feedback from diverse audiences into iterative improvements [3, 42, 72]. Over the past decade, advances in Human-Computer Interaction (HCI) have enabled writers to easily obtain feedback from instructors, peers, online communities [25, 65], crowdsourcing platforms [51, 77], and generative AI [4]. These tools enhance feedback availability, scaling it to dozens—or even hundreds—of free-form textual responses.

However, easy access to feedback can be a double-edged sword. Feedback from diverse audiences may contain contradictions, focus on different topics, and vary widely in structure. This makes it hard to find emerging patterns, reconcile conflicting ideas, and prioritize revisions [79]. For example, when an argumentative essay is critiqued, a reviewer might point out the need for stronger evidence, suggest clarifying the thesis statement, and recommend addressing potential counterarguments—all within the same textual comment. With additional reviewers, the volume of critiques can expand, covering elements such as claims, warrants, evidence, and rebuttals. More reviewers also creates more viewpoint diversity, expanding the complexity of potential changes.

To act upon feedback effectively, writers must distill lengthy responses into actionable insights, prioritize high-value suggestions, pinpoint problematic areas, and revise their work to incorporate selected insights [3, 18, 42, 72, 79]. This process is **iterative** and **non-linear**: writers often assess how different suggestions may impact their work, experiment with various revision strategies, and ultimately determine whether and how each comment should be addressed [66]. Yet, the **scalability** and **variability** within a feedback set can leave writers navigating long blocks of unstructured comments, feeling uncertain about where to begin and struggling to prioritize, trace, or flexibly act on diverse suggestions [18, 31].

Existing tools focus primarily on revision generation, such as suggesting alternative phrasings or enabling one-shot rewrites using language models [1, 34, 48, 57, 66]. While useful, such tools often flatten feedback into summaries, obscure underlying reasoning, or assume a linear revision path. Few systems support this kind of exploratory feedback use or help writers interpret, organize, and experiment with suggestions in ways that preserve agency and enable transparent, strategic decision-making.

In this paper, we introduce Synthia, a system designed to support **sensemaking in feedback-driven writing revision**. Grounded in established best practices for feedback-based revision, Synthia advances prior work through three key innovations:

(1) it offers configurable visual encodings of feedback (i.e., bubbles) that preserve the granularity of individual suggestions, enabling writers to surface patterns, assess helpfulness, and prioritize critiques based on evolving goals;
(2) it introduces bidirectional links between feedback, source text, and revisions, allowing writers to trace the context, relevance, and impact of specific comments; and
(3) it supports non-linear revision through branching paths and iterative exploration, encouraging experimentation rather than one-shot rewriting.

To validate our design, we conducted a within-subjects user study ($N = 12$) comparing Synthia with a baseline writing interface featuring a GPT-4-based chat assistant and version tracking. Our findings revealed that participants using Synthia identified more justified and actionable feedback comments, developed more strategic and exploratory revision paths, and reported greater ownership and transparency in their process. These findings highlight how innovations in Synthia can support writers in making sense of feedback and iteratively improving their work. We conclude with implications for designing revision tools that scaffold not just rewriting, but the interpretation, navigation, and application of feedback in all its complexity.

2 Related Work

2.1 Interactive Feedback Tools

HCI researchers have developed a range of tools to assist creators in efficiently collecting high-quality feedback at scale [4, 10, 24, 39, 51, 76, 81]. As people from diverse backgrounds and areas of expertise may prioritize issues differently, reviewers can offer varying, sometimes contradictory, opinions on the same content [28]. This complexity of feedback makes its effectiveness dependent on recipients' ability to interpret, learn, and act on it [3, 18, 19, 42, 72, 79].

First, prioritizing feedback requires creators to assess its helpfulness. Guo et al. [25] highlight that feedback's sentiment, actionability, justification, and specificity correlate with recipients' willingness to invest effort in improving their artifacts. However, existing feedback tools often emphasize surface-level attributes such as topic and sentiment, leaving other informative qualities implicit. We aim to encode these deeper attributes into visualizations to help users better discern the value of each comment.

Second, navigating large volumes of feedback is challenging. To support this, HCI researchers have developed tools that help users interpret and organize feedback [11, 31, 51, 76, 79]. For example, Decipher [11, 79] shows topic distributions across providers in tabular format. OpinionSpace [16] maps community feedback into a 2D space, using color to indicate sentiment and point size to reflect community endorsement. While these visualizations help surface patterns, they are predefined by tool designers. In contrast, we aim to give writers control over how feedback is visualized—allowing them to choose which attributes to include based on their goals.

Finally, identifying areas for revision and mapping feedback to corresponding text segments is essential for action planning, but

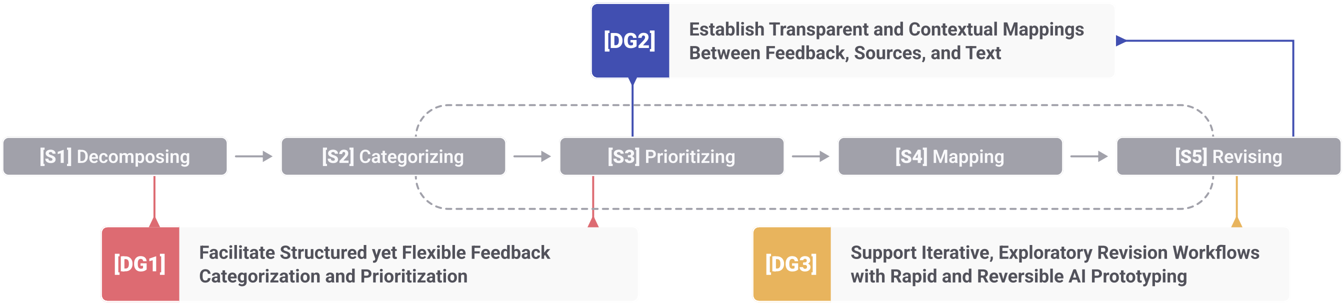

[DG2] Establish Transparent and Contextual Mappings Between Feedback, Sources, and Text
[S1] Decomposing
[S2] Categorizing
[S3] Prioritizing
[S4] Mapping
[S5] Revising
[DG1] Facilitate Structured yet Flexible Feedback Categorization and Prioritization
[DG3] Support Iterative, Exploratory Revision Workflows with Rapid and Reversible AI Prototyping

Figure 2: Feedback-based writing revision process with three design goals for Synthia. We identified five core practices from the literature [18, 19, 40, 43, 66, 79] and associate them with three specific design goals for Synthia (described in §3).

is cognitively demanding [18]. Most textual feedback lacks clear indications of which parts of the writing need attention. Writers must also weigh their available time and expertise when approaching revisions. While existing tools surface feedback patterns, they often lack support for helping users gauge the scope and location of necessary changes based on selected comments. Our tool addresses this gap by highlighting potential revision areas, helping users better assess the impact of selected feedback.

Our work advances feedback tools by supporting the entire feedback cycle—from exploration and interpretation to implementation. We leverage the power of LLMs to enable rapid prototyping, combined with interactive visual scaffolding that empowers users to take greater control over their revision goals.

2.2 Intelligent Writing Tools

The HCI community has a long-standing interest in designing writing tools [46]. Projects support writers across various stages, such as brainstorming ideas [21, 59], planning outlines [85], drafting content [9, 14, 30, 35, 41, 80], and refining text [1, 34, 48, 57, 66].

Among revision tools, commercial applications like Grammarly and Ref-N-Write [50] focus on addressing conventions, grammar, and stylistic improvements. Academic tools extend this by offering fluent sentence alternatives [15, 32–34, 37, 53], often presenting multiple options with scores to guide writers [6, 17, 20, 32, 47]. Another line of tools provide visual, numerical, or textual assessments of writing quality to guide revision [54, 75]. For instance, AL [70] visualizes argumentative sentence relationships and scores persuasiveness, while ArgRewrite [84] tracks and assesses sentence-level changes. In these tools, sentences represent natural textual boundaries, allowing clear demarcation of edits and facilitating easy tracking and application of changes. Building on these interfaces, we also adopt sentences as the scope of interactive revision spans.

HCI and traditional design practices encourage parallel exploration of multiple variations to help creators avoid fixation on a single idea [23, 36]. Similarly, experienced writers approach revision as a recursive, non-linear process [66, 71] and engage with the text in repeated cycles. Inspired by this philosophy, Reza et al. [57] introduced ABScribe, which facilitates rapid exploration of multiple writing variations through LLM-based human-AI co-writing. It showed the potential of LLMs to quickly generate new versions of writing pieces. While generated text may lack the fidelity required for final drafts, imperfect AI text can help writers rapidly explore revision possibilities [57, 80], which inspired us to have it serve as a lens for deepening user understanding of feedback.

Prior work provides limited support for helping writers integrate feedback into revisions. One exception is Impressona [4]. It generates feedback based on writer-defined AI personas representing target readers. This approach further lowers barriers to obtaining diverse critiques. However, there is still a gap between receiving and implementing feedback. To support feedback-based revision, we introduce a novel tool that visualizes and synthesizes feedback with LLM-generated revisions, enabling writers to critically evaluate, experiment with, and refine their work through iterative cycles of feedback integration.

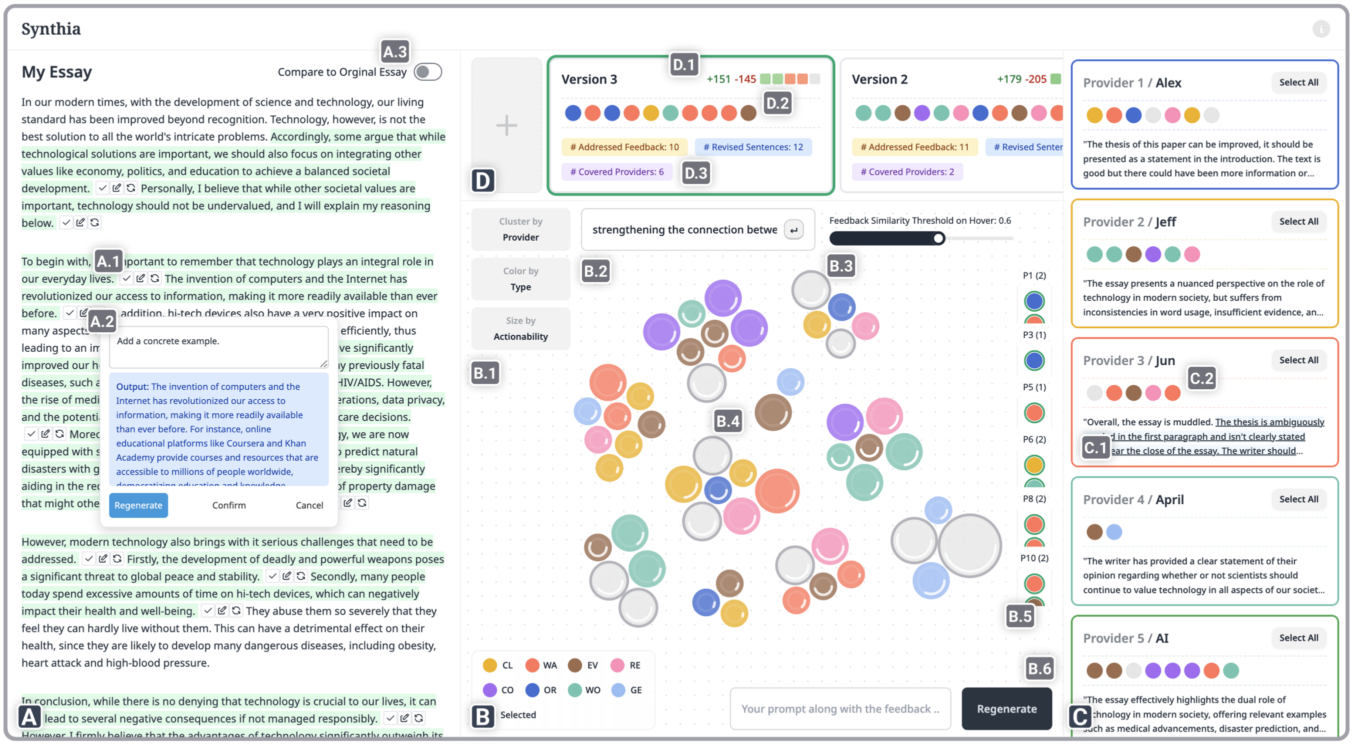
2.3 Visual Interfaces for LLMs

In recent years, LLMs have reshaped how we acquire, process, and interact with information. However, the linear, text-heavy nature of traditional conversational user interfaces (CUIs) has been criticized for hindering user sensemaking of LLM-generated content. To address this issue, HCI researchers develop various visual interfaces to scaffold sensemaking. For example, Luminate [63] structures dimensional reasoning to help writers explore design spaces. Sensecape [64] and Graphologue [38] use interactive mind maps to streamline information foraging. Other work, such as text rendering techniques from Gero et al. [22], supports mesoscale (10s to 100s of samples) sensemaking of LLM responses.

While these efforts focus on parsing and organizing outputs, a critical gap remains: when dozens of feedback comments require curation, combination, and experimentation, users lack tools to manage this complexity. Moreover, few systems reveal relationships between information sources, such as contextualizing feedback in relation to the essay's content (input↔input) or how it translates into revisions (input↔output). Without transparent tools to expose these relationships, users struggle to exchange appropriate information with LLMs and trace how inputs (feedback) influence outputs (revisions). Our work addresses this gap by designing an interface for managing mesoscale feedback, enabling writers to manage comments and trace implications.

2.4 Best Practices for Feedback-Based Revision

Finally, we survey literature on current best practices for feedback-based revision so that we can inform eventual design goals for our



My Essay

In our modern times, with the development of science and technology, our living standard has been improved beyond recognition. Technology, however, is not the best solution to all the world's intricate problems. Accordingly, some argue that while technological solutions are important, we should also focus on integrating other values like economy, politics, and education to achieve a balanced societal development. ✓ Personally, I believe that while other societal values are important, technology should not be undervalued, and I will explain my reasoning below. ✓

To begin with, ...portant to remember that technology plays an integral role in our everyday lives. ✓ The invention of computers and the Internet has revolutionized our access to information, making it more readily available than ever before. ✓ ... addition, hi-tech devices also have a very positive impact on
many aspects ... efficiently, thus
leading to an im... ...ve significantly
improved our h... ...y previously fatal
diseases, such a... ...HIV/AIDS. However,
the rise of medi... ...erations, data privacy,
and the potenti... ...care decisions.
✓ Morec... ...gy, we are now
equipped with s... ...o predict natural
disasters with g... ...ereby significantly
aiding in the rec... ...of property damage
that might othe...

Add a concrete example.

Output: The invention of computers and the Internet has revolutionized our access to information, making it more readily available than ever before. For instance, online educational platforms like Coursera and Khan Academy provide courses and resources that are accessible to millions of people worldwide, democratizing education and knowledge

However, modern technology also brings with it serious challenges that need to be addressed. ✓ Firstly, the development of deadly and powerful weapons poses a significant threat to global peace and stability. ✓ Secondly, many people today spend excessive amounts of time on hi-tech devices, which can negatively impact their health and well-being. ✓ They abuse them so severely that they feel they can hardly live without them. This can have a detrimental effect on their health, since they are likely to develop many dangerous diseases, including obesity, heart attack and high-blood pressure.

In conclusion, while there is no denying that technology is crucial to our lives, it can lead to several negative consequences if not managed responsibly. ✓
However, I firmly believe that the advantages of technology significantly outweigh its

Version 3 +151 -145
Addressed Feedback: 10
Revised Sentences: 12
Covered Providers: 6

Version 2 +179 -205
Addressed Feedback: 11
Revised Senter
Covered Providers: 2

strengthening the connection betwe ↵

Your prompt along with the feedback ..

Provider 1 / Alex

"The thesis of this paper can be improved, it should be presented as a statement in the introduction. The text is good but there could have been more information or...

Provider 2 / Jeff

"The essay presents a nuanced perspective on the role of technology in modern society, but suffers from inconsistencies in word usage, insufficient evidence, an...

Provider 3 / Jun

"Overall, the essay is muddled. The thesis is ambiguously ...d in the first paragraph and isn't clearly stated ...ear the close of the essay. The writer should...

Provider 4 / April

"The writer has provided a clear statement of their opinion regarding whether or not scientists should continue to value technology in all aspects of our societ...

Provider 5 / AI

"The essay effectively highlights the dual role of ...chnology in modern society, offering relevant examples ...uch as medical advancements, disaster prediction, and...

Figure 3: User Interface of Synthia. The interface consists of four main components: (A) the Essay Panel, where users review their original essays and refine their revisions; (B) the Bubble Canvas, which allows users to categorize, prioritize, and select feedback for revision; (C) the Feedback Gallery, where users can explore detailed feedback from different providers; and (D) the Revision Panel, which tracks multiple versions of the draft and their respective changes.

system. Prior work broadly considers: creators’ strategies and practices in managing and interpreting feedback [19, 25, 40, 43, 79]; the role and nuances of revision within the writing process [18, 29, 66, 71]; and the latest research on writing and feedback tools [31, 46, 57, 79, 80, 85]. Drawing from this body of work, we identify five core practices for effective feedback-based revision: decomposing, categorizing, prioritizing, mapping, and revising (see Fig. 2).

Writers start by **decomposing [S1]** lengthy responses into manageable units, identifying key critiques while fighting cognitive fatigue that comes from sustained parsing efforts [18, 79]. This prepares the ground for **categorization [S2]**, where fragmented comments are organized according to providers, purposes, or emotions to develop a high-level model of issues [79].

Given time constraints and conflicting perspectives, writers cannot address everything; they must **prioritize [S3]** feedback that aligns with their rhetorical goals (e.g., strengthening claims, adding examples) and that offers reasonable, actionable, and specific suggestions [43]. However, this task is complicated by the implicit nature of value judgments in feedback language, which makes it hard for recipients to identify high-impact opportunities [73]. Then, they go deep in **mapping [S4]** feedback: writers synthesize scattered yet related voices and trace critiques to textual targets [18]. This phase bridges feedback and text, requiring writers to maintain a mental map that links feedback to corresponding passages.

Lastly, writers engage in **revision [S5]** as an iterative dialogue. Rapid prototyping of changes helps concretize abstract suggestions and reveal hidden trade-offs between competing values [42, 66]. This final stage demands advanced writing skills to evaluate multiple potential revisions while managing the cognitive complexity of the entire feedback cycle.

3 System Design and Implementation

We introduce Synthia, an interactive system that supports writers in interpreting and synthesizing feedback into writing revisions by: (1) reifying feedback as interactive, configurable bubbles, (2) surfacing contextual, cross-information relationships, and (3) prototyping in parallel with accessible, traceable drafting.

We demonstrate our system with respect to argumentative writing, a domain requiring writers to present and defend their perspective on a specific topic. To effectively convey their argument, writers must engage in clear reasoning, consider alternative viewpoints, and refine their stance into a persuasive written piece. This process offers a valuable opportunity for feedback, as writers must engage with other perspectives to improve their work [44, 45, 56]. While we focus on argumentation, our approach generalizes to other writing genres involving feedback-driven revision.

The interface of Synthia consists of four main sections: Essay Panel (Fig. 3A), Bubble Canvas (Fig. 3B), Feedback Gallery (Fig. 3C), and Revision Panel (Fig. 3D). Below, we present each of the three

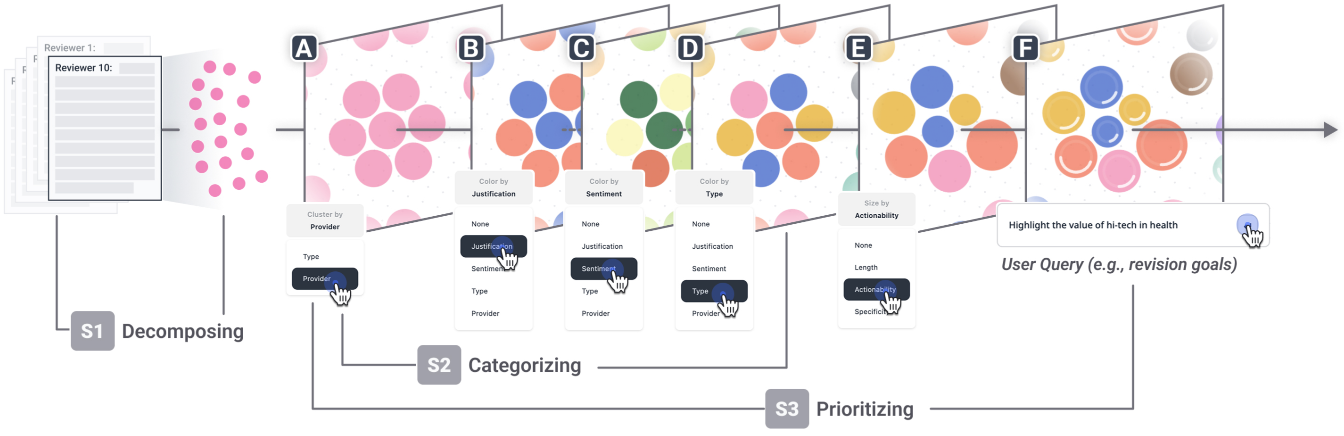


Figure 4: Feedback units are represented as interactive, configurable bubbles in Synthia. Users can cluster these bubbles by type or provider (A), color them based on justification (B), sentiment (C), provider, or type (D), and resize them according to feedback length, actionability, or specificity (E). Additionally, users can visualize the relevance of each feedback unit to a custom query through the radian of reflection arcs on each bubble (F).

design goals (Fig. 2) alongside the corresponding core interaction in Synthia that was informed by it. Then, we demonstrate its workflow by illustrating how an imagined user, Choi, would interact with the system to fulfill her revision goals. Lastly, we provide a concise overview of the system implementation.

3.1 Reifying Feedback as Interactive, Polymorphic Bubbles [DG1]

To support decomposition [S1], categorization [S2], and prioritization [S3] of feedback, the system should first automate the extraction of discrete feedback idea units and surface implicit attributes (e.g., purpose, sentiment, helpfulness). To reduce cognitive load, it should also provide visual scaffolds for filtering, grouping, and weighting these units, while enabling dynamic switching between different organizational models. This balance of structure and flexibility will help writers navigate feedback complexity without imposing rigid workflows [76, 79]. This leads us to our first design goal: **[DG1]** *Facilitate Structured yet Flexible Feedback Categorization and Prioritization.*

To achieve this aim, our system automatically breaks uploaded lengthy reviews into smaller, individual feedback units, each defined as one or more sentences that express a coherent critique [79]. Each unit is represented as an interactive, configurable bubble (●) (Fig. 3B.4). Users can configure various visual encodings, such as location, color, size, and arcs, based on the feedback unit's categorical and numerical attributes to reveal distribution patterns and assess the helpfulness.

Synthia encodes two **categorical attributes**: **`Provider`** (source of the feedback unit) and **`Type`** (writing issue addressed by the feedback unit), which can be used to cluster or color the feedback bubbles (Fig. 3B.1). For types, each feedback unit is assigned to one of the eight writing issue categories proposed [83]: Claims / Ideas (●), Warrant / Reasoning / Backing (●), Evidence (●), Rebuttal / Reservation (●), Convention / Grammar / Spelling (●), Word-Usage / Clarity (●), Organization (●), and General Content (●). We follow prior work to code surface-level issues (e.g., grammar, word-usage) in cold colors (e.g., blue) and content-level issues (e.g., claims, evidence) in warm colors (e.g., orange) [1, 84].

The system encodes five helpfulness metrics drawn from prior work [7, 25, 43] as **numerical attributes**:

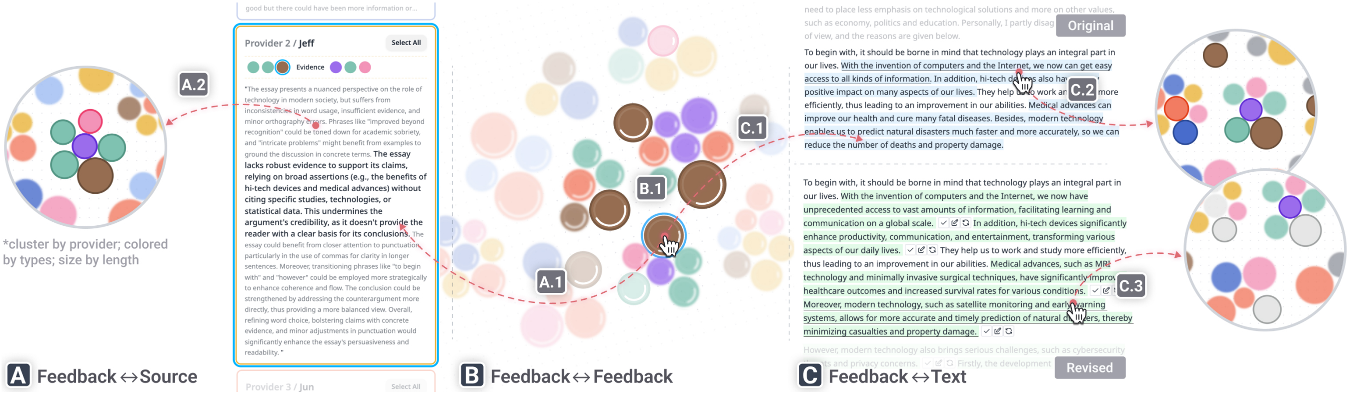
- **`Justification`** (binary): Whether the feedback unit is justified with explanations.
- **`Sentiment`** (sequential): Valence of the feedback unit (ranging from negative to positive).
- **`Actionability`** (sequential): Number of actionable suggestions provided in the feedback unit.
- **`Specificity`** (sequential): Level of detail in the feedback.
- **`Length`** (sequential): Number of words in the feedback.

With these numerical attributes, users can color bubbles by justification (i.e., ● as justified, ● as unjustified; as shown in Fig. 4B), sentiment (i.e., gradient color from red/negative to green/positive; as shown in Fig. 4C), or resize bubbles based on length, actionability (Fig. 4E), or specificity to evaluate feedback helpfulness.

Moreover, to help users prioritize feedback at the content level, the system encodes the radian of reflection arcs to represent the relevance of each feedback unit to the user query. Users can enter their revision goals in the search bar (Fig. 3B.2). Synthia will compute the embedding similarity (ranging from 0 to 1) between the feedback unit and the query and then highlights this similarity by drawing a reflection arc inside the corresponding bubble (Fig. 4F).

3.2 Surfacing Contextual, Cross-Information Connections [DG2]

For effective feedback mapping [S4], the system should surface three core relationships: First, it should trace critiques to their specific reviewers, contextualizing comments to clarify intent (feedback ↔ source). Second, it should highlight similar voices from available critiques, helping writers identify patterns (feedback ↔ feedback). Third, the system should anchor critiques to specific sections of text and track how edits respond to feedback, allowing



good but there could have been more information or...

Provider 2 / Jeff Select All

Evidence

"The essay presents a nuanced perspective on the role of technology in modern society, but suffers from inconsistencies in word usage, insufficient evidence, and minor orthography errors. Phrases like "improved beyond recognition" could be toned down for academic sobriety, and "intricate problems" might benefit from examples to ground the discussion in concrete terms. **The essay lacks robust evidence to support its claims, relying on broad assertions (e.g., the benefits of hi-tech devices and medical advances) without citing specific studies, technologies, or statistical data. This undermines the argument's credibility, as it doesn't provide the reader with a clear basis for its conclusions.** The essay could benefit from closer attention to punctuation, particularly in the use of commas for clarity in longer sentences. Moreover, transitioning phrases like "to begin with" and "however" could be employed more strategically to enhance coherence and flow. The conclusion could be strengthened by addressing the counterargument more directly, thus providing a more balanced view. Overall, refining word choice, bolstering claims with concrete evidence, and minor adjustments in punctuation would significantly enhance the essay's persuasiveness and readability. "

Provider 3 / Jun Select All

need to place less emphasis on technological solutions and more on other values, such as economy, politics and education. Personally, I partly disag of view, and the reasons are given below.

To begin with, it should be borne in mind that technology plays an integral part in our lives. With the invention of computers and the Internet, we now can get easy access to all kinds of information. In addition, hi-tech de s also hav positive impact on many aspects of our lives. They help o work an more efficiently, thus leading to an improvement in our abilities. Medical advances can improve our health and cure many fatal diseases. Besides, modern technology enables us to predict natural disasters much faster and more accurately, so we can reduce the number of deaths and property damage.

To begin with, it should be borne in mind that technology plays an integral part in our lives. With the invention of computers and the Internet, we now have unprecedented access to vast amounts of information, facilitating learning and communication on a global scale. ✓ In addition, hi-tech devices significantly enhance productivity, communication, and entertainment, transforming various aspects of our daily lives. ✓ They help us to work and study more efficiently, thus leading to an improvement in our abilities. Medical advances, such as MRI technology and minimally invasive surgical techniques, have significantly improv healthcare outcomes and increased survival rates for various conditions. ✓ Moreover, modern technology, such as satellite monitoring and early warning systems, allows for more accurate and timely prediction of natural d ers, thereby minimizing casualties and property damage. ✓

However, modern technology also brings serious challenges, such as cybersecurity ts and privacy concerns. ✓ Firstly, the development

Figure 5: SYNTHIA visualizes three key relationships through bidirectional highlighting: (A) between a feedback unit and its original context, (B) between a feedback unit and similar feedback from other providers, and (C) between a feedback unit and the sentences it targets or impacts. These highlights are triggered when users hover over a feedback bubble (A.1 & B.1 & C.1), hover over a feedback card (A.2), or click on a sentence (C.2 & C.3).

writers to link feedback with corresponding text and evaluate revision impact (feedback ↔ text). By making these connections explicit, the system bridges the gap between critique and actionable revision. Our next design goal considers these mappings: **[DG2]** *Establish Transparent and Contextual Mappings Between Feedback, Sources, and Text.*

To accomplish this goal, SYNTHIA visualizes three key relationships via bidirectional highlighting: the feedback unit and its context, the feedback unit and its similar counterparts, and the feedback unit and its targeting and impacted sentences.

Feedback↔Source: Each reviewer's complete comment is initially collapsed into a summary card in Feedback Gallery (Fig. 3C). Hovering over a feedback bubble expands its source card, highlighting the original comment within its broader context (Fig. 5A.1). Conversely, hovering a feedback card highlights all associated bubbles in the canvas (Fig. 5A.2). Users can batch-select all feedback bubbles of this provider or exclude them for revision by clicking the `Select/Deselect All` button. This bidirectional binding ensures users never lose the contextual grounding between abstract visual encodings (bubbles) and their textual origins.

Feedback↔Feedback: When users hover over a feedback bubble, SYNTHIA dynamically surfaces semantically similar critiques. The system calculates embedding similarity between the hovered unit and all others. This approach follows established practices in semantic textual similarity, where embeddings have been shown to effectively capture nuanced semantic relationships in text [49, 67]. Similar voices are highlighted: the hovered bubble gains a blue outline, while its counterparts are emphasized with thicker, darker borders and brought to the foreground to avoid overlap (Fig. 5B.1). Non-matching bubbles are rendered with a Gaussian blur effect to reduce visual clutter. To curate groups, users can hold `Shift` and click to select both the hovered bubble and its matches, enabling batch operations like collective prioritization or exclusion.

Feedback↔Text: To map abstract feedback to concrete text segments, SYNTHIA predicts the relevant sentences that need revisions for each feedback unit. With this prediction, when users hover over a bubble, the system highlights potentially problematic sentences it targets (Fig. 5C.1). Similarly, holding `Shift` and clicking a sentence reveals all associated feedback, letting users batch-select critiques for a specific sentence or exclude irrelevant ones (Fig. 5C.2). Post-revision, edited sentences retain visual links to their original feedback, enabling writers to audit how revisions addressed associated critiques or introduced new issues (Fig. 5C.3).

3.3 Prototyping in Parallel with Accessible, Traceable Drafting [DG3]

Lastly, to support revision [S5], the system should facilitate nonlinear workflows that include rapid prototyping, version comparison, and reversible drafting. LLM-generated revisions, though imperfect, can serve as scaffolding to help writers concretize feedback and test hypotheses. To this end, it should: (1) provide easy-to-access AI drafting, (2) support version tracking and comparison, (3) require explicit user approval before integrating suggestions, and (4) enable manual adjustments via direct editing or regeneration. Summarized as a design goal: **[DG3]** *Support Iterative, Exploratory Revision Workflows with Rapid and Reversible AI Prototyping.*

SYNTHIA encourages writers to rapidly prototype. Users begin by selecting feedback bubbles which populate the Preparation Station (Fig. 3B.5). Next, users can click the `Generate` button (Fig. 3B.6) and optionally provide additional instructions in a text box. It will prompt the system to revise the associated sentences based on the chosen feedback. This draft serves as a concrete starting point, allowing users to assess how critiques might reshape their text while maintaining editorial control. If unsatisfied, users can click the `Regenerate` button (Fig.3B.6) to get a new one. Alternatively, each revised sentence is accompanied by three small action icons (Fig. 3A.1). Users can `Accept` (✓) the suggestion, `Edit` (✎) manually to directly modify the proposal, or `Regenerate` (↻) to refine via follow-up prompts (Fig. 3A.2).

Synthia supports non-linear, iterative, and traceable experimentations. Users can branch into different revision paths by applying different feedback combinations. The Revision Panel tracks all versions, enabling side-by-side comparisons with metrics such as word changes (Fig. 3D.1), addressed feedback count (Fig. 3D.2), revised sentence count, and contributor diversity (Fig. 3D.3). Additionally, users can enable a comparison mode (Fig. 3A.2) to visualize differences between a selected version and the original text.

3.4 Example Scenario

Choi, a university student, is working on an essay about whether society should place less emphasis on technological solutions and focus more on other values. After completing her draft, she receives feedback from her teacher, classmates, an AI writing assistant, and the EssayForum[1]. Faced with ten detailed reviews, each consisting of one to three paragraphs, Choi finds it overwhelming so she uploads her essay and the collected reviews into Synthia for assistance.

Synthia decomposes the reviews into 54 feedback bubbles. Choi first clusters by provider to assess distribution and then switches to a type-based view to identify overarching issues, revealing that most critiques center on claim, reasoning, and evidence. To prioritize, she colors the bubbles by justification and adjusts the size encoding to highlight actionable feedback (Fig. 3B.1). She also enters her revision goal, "*strengthening the connection between claims and evidence*," which prompts Synthia to highlight relevant feedback through arc indicators (Fig. 3B.2).

Hovering over a critique about evidence on healthcare surfaces three similar comments, helping her recognize a recurring issue. She batch-selects these related feedback units, which then highlights the corresponding sentence in her essay (Fig. 5C.1). Curious about another section on evidence, she clicks a sentence (Fig. 5C.2), revealing linked critiques; she removes redundant ones and keeps ones addressing evidence specificity. After finalizing units of feedback, Choi then clicks `Generate` to synthesize these comments into a new version. To evaluate the generated revisions, she clicks a revised sentence to highlight the corresponding critiques (Fig. 5C.3). She accepts some revised sentences while manually refining others to maintain her writing style. Dissatisfied with an overly technical example, she iterates on alternatives (Fig. 3A.2).

To explore different revision strategies, Choi branches two more parallel versions: one refining reasoning by defining "*technological solutions*" with renewable energy examples, and another strengthening evidence with low-tech public health case studies. Comparing these versions in the Revision Panel, she finalizes a hybrid draft but notices new "*redundant phrasing*" flags. Hovering these issued sentences links them to word-usage feedback; she regenerates alternatives and exports her essay. Through this workflow, Synthia helps Choi efficiently prioritize, interpret, and act upon feedback, ultimately producing a stronger argumentative essay.

3.5 Implementation Notes

The frontend of Synthia is built on Next.js, leveraging server-side rendering to manage API calls to to Firebase for event logging and OpenAI for LLM access. Interactive bubble visualizations use the force-directed layout algorithm from d3.js, featuring collision detection and charge forces optimized for spatial clustering.

The helpfulness metrics for feedback (§3.1) are computed using a series of computational linguistic pipelines proposed by Krause et al. [7, 43]. They are implemented using the pattern.en package and the Natural Language Toolkit in a Python Flask backend. The text embedding for semantic similarity calculations is obtained via the `text-embedding-3-small` model from OpenAI. In addition, we prompt `GPT-4o` to break down feedback, classify feedback purposes (§3.1), predict problematic sentences (§3.2), and draft or regenerate revisions (§3.3), based on the prompting strategies proposed by Wu et al. [74]. Prompts, few-shot examples, and sample outputs are available in supplementary materials.

4 Technical Evaluation

To iterate our prompts and assess the validity of our LLM pipelines prior to use by human participants, we conducted a technical evaluation on Synthia's ability to (1) accurately classify feedback units, (2) detect relevant sentences based on feedback, and (3) generate improved sentence revisions.

We first constructed a feedback corpus through human annotation to establish ground truth. This corpus was built on six writing samples from an established argumentative essay dataset [62]. These essays, collected from EssayForum, cover diverse topics including education, technology, and economic policy. For each writing sample, we recruited ten crowd workers from Prolific to provide feedback guided by a rubric [26]. Prior research demonstrated that crowdsourced feedback with rubrics achieves quality, scope, and depth comparable to expert or community critiques [77, 78, 81]. To mitigate LLM use, we followed the safeguards from the work of Veselovsky et al. [68]: explicit instructions prohibiting LLM assistance and conversion of the textual rubric into an image. Workers were compensated with $4 per task. A research assistant segmented the 60 feedback entries into 223 discrete feedback items, each addressing a single issue in the target essays. We randomly sampled 100 items for annotation and downstream evaluation.

All human evaluations were conducted by three research assistants who are proficient in English. Each has years of experience in academic writing and has completed two semesters of specialized training in argumentative writing skills. Before beginning each task, evaluators received further training until their inter-rater reliability reached a satisfactory level on trial tasks.

4.1 Performance of Feedback Classification

Two research assistants annotated the types of 100 sampled feedback items in the dataset (Cohen's $\kappa = 0.80$). After resolving 13 discrepancies through discussion, the final ground-truth distribution comprised 15% Claim, 20% Warrant, 26% Evidence, 7% Rebuttal, 14% Conventions, 3% Word-usage, 12% Organization, and 3% General Content/Others feedback units. Our prompted LLM achieved an overall precision of 0.90, recall of 0.84, and macro F1-score of 0.84. A closer analysis showed that although several unique or rare units about General Content/Others (33% accuracy) were misclassified, the model performed robustly in other categories: 100% in

[1] EssayForum.com is a non-profit online community for writers to solicit feedback on their essays.

Word-usage, Evidence, and Rebuttal, 93% in Conventions, 87% in Claim, 83% in Organization, and 75% in Warrant.

4.2 Performance of Sentence Detection

Given feedback samples and corresponding essays, the LLM pipeline identified an average of 2.6 problematic sentences per sample. Identifying problematic sentences based on open-ended feedback is an inherently subjective and interpretive task, with no single correct answer or established ground truth. Thus, we assessed performance via expert ratings. Three evaluators rated the relevance of detected sentences to the feedback on a 5-point Likert scale (1: irrelevant, 5: highly relevant), achieving an inter-rater reliability of 0.84. Results indicated high performance ($M = 4.41$, $SD = 0.60$).

Evaluators commented that our LLM method effectively scanned the entire essay for issues, not only catching problematic sentences in noticeable areas like the beginning paragraph but also thoroughly reviewing the body paragraphs. However, the LLM sometimes failed to identify all relevant sentences when the feedback was overly specific. For instance, when feedback highlighted specific sentences that needed revision (e.g., "*revise the third sentence*"), the LLM identified only the named sentence but failed to recognize additional related sentences elsewhere in the essay.

4.3 Performance of Sentence Revision

To evaluate revision quality, we fine-tuned GPT-3.5 using Afrin and Litman's Revision Quality dataset [2], which labels 940 original-revised sentence pairs as "better" or "not better." We divided the dataset into training (60%), validation (20%), and test sets (20%). The model was fine-tuned on the training and validation sets, and finally achieved high levels of precision (0.89), recall (0.86), and a macro F1-score (0.87) on the test set. Applying this model to our LLM-generated revisions showed that 100% revised sentences were classified as "better" than their originals.

However, since syntactic improvement does not guarantee feedback resolution, we further tasked two human evaluators to manually verify whether the revisions addressed the targeted feedback (Cohen's $\kappa = 0.78$). Results revealed that 84% of revisions resolved the feedback intent, with failures primarily occurring when feedback required structural or content reorganization. Evaluators further commented that the LLM excelled at creative revision tasks like adding evidence, strengthening warrants, or crafting rebuttals. For example, it may inject domain-specific examples to bolster claims, transform vague claims into cause-effect chains, and anticipate counterarguments from different perspectives. These findings suggest that our LLM pipeline can not only improve syntactic quality but also meaningfully integrate feedback to enhance argumentative depth, highlighting its potential to help users quickly explore revision ideas by synthesizing multiple feedback cues.

5 User Evaluation

To further evaluate the efficacy of SYNTHIA, we conducted a within-subjects study[2] where we compared the system to a carefully constructed baseline. We aim to address four research questions:

RQ1 *How does SYNTHIA's interactive feedback visualization help writers prioritize, interpret, and act on feedback?*

RQ2 *How does SYNTHIA help writers translating feedback into revisions, particularly in exploring different revision pathways?*

RQ3 *How do writers perceive the AI support provided by SYNTHIA in feedback-driven revision?*

RQ4 *What challenges do users encounter when using SYNTHIA in feedback-driven revision?*

5.1 Baseline

In the Baseline[3] condition, we preserved the same overall UI layout as SYNTHIA and allowed users to save and track versions of their drafts. The key difference was that SYNTHIA's interactive feedback visualization was replaced with a conversational AI assistant powered by the same underlying LLM, consistent with approaches used in prior work (e.g., [57]). To enable seamless AI interaction, the original essay text was embedded directly into the system prompt of the assistant as context. Users could specify areas for revision and select feedback by clicking relevant sections, removing the need to manually copy and paste text. This approach 1) maintains basic SYNTHIA features (e.g., importing data, app-like appearance) that do not relate to the contributions, thereby reducing interface confounds, and 2) mirrors real-world practices where users have access to general-purpose AI tools like ChatGPT. Baseline also represents a middle ground between traditional manual revision and fully automated feedback processing and rewriting. Users had open access to the full capacity of a general-purpose AI tool but were not required to use it in any particular way. They could request feedback summaries, generate end-to-end automated revisions from all or selected feedback, or engage the assistant with custom prompts, supporting both partial and full automation workflows.

5.2 Participants

We recruited 12 participants (8 female, 4 male) aged 20—36 ($M = 24$, $SD = 4.37$) from a private R1 university in the United States. The group included 2 sophomores, 2 juniors, 2 seniors, and 6 graduate students. All participants were proficient in English, and 6 were native English speakers. When asking about their argumentative writing skills, 5 participants identified themselves as intermediate writers (having some experience but still honing their skills), 3 as advanced writers (with significant experience and confidence in their abilities), and 4 as an expert writer (having extensive experience and a high level of skill). In addition, the average incoming freshman at this institution scores in the 99th percentile nationally on the SAT Evidence-Based Reading and Writing (EBRW) section, and all students complete two semesters Writing & Rhetoric training in their first year. Thus, we anticipated that most, if not all, of our participants would be proficient writers. Additionally, all participants reported that they often used generative AI tools, especially ChatGPT, in their daily writing practice.

5.3 Task Materials

Participants received two essays from the aforementioned dataset [62]. Both essays were comparable in scope—one discussing technology development and the other addressing mobile phones—with each

[2] The study received approval from our institution's IRB.

[3] The Baseline user interface is provided in the supplementary materials.

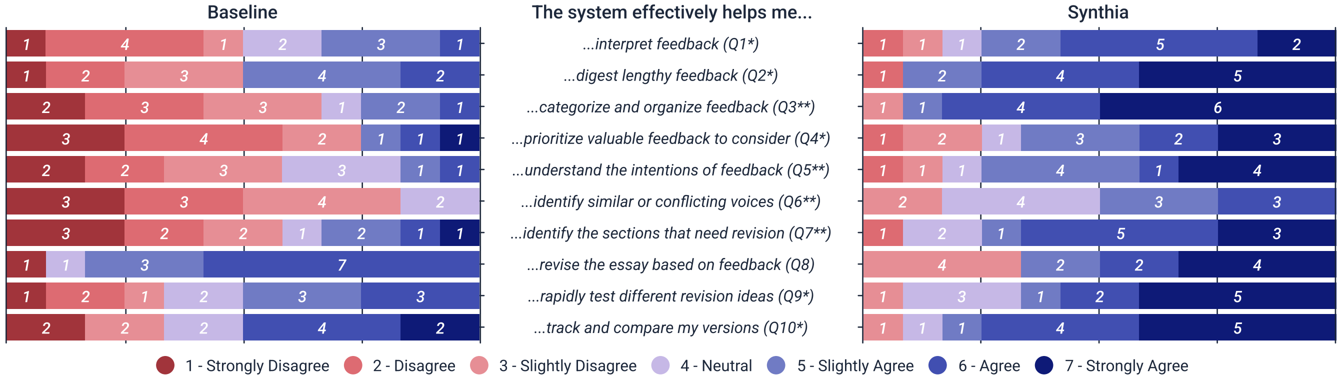


Figure 6: Participants' responses to a 7-point self-defined Likert scale questionnaire, measuring their perceived support across various dimensions of feedback-based revision in both the Baseline condition and our system.

approximately 300 words long. Following the method outlined in §4, we collected 10 feedback responses per essay through crowdsourcing, totaling approximately 2,000 words per set. Three evaluators from the technical evaluation rated the feedback's perceived usefulness using a 7-point Likert scale ($ICC = 0.77$). A Mann-Whitney U test revealed no statistically significant difference between the feedback sets (5.2 *vs.* 5.3; $W = 53$, $p = .848$).

5.4 Study Procedure

The study began with researchers obtaining informed consent and collecting demographic information. Participants then engaged in two separate task sessions, each beginning with a 3–5 minute tutorial followed by a 20-minute feedback-based revision task using either Synthia or the Baseline system. The task materials and system conditions were counterbalanced.

The tutorial introduced the key features of the assigned tool. Since all participants had prior experience with AI writing tools, we refrained from mentioning specific prompts to avoid influencing their natural approach. After the tutorial, we asked participants to imagine a scenario where their essay received ten pieces of feedback from their peers and online writing community members. They were then required to revise their essay[4] based on these ten feedback responses using the assigned tool. Participants were encouraged to explore different versions during revision based on the provided feedback. At the end of each task session, they selected one version as their final submission.

After each task session, participants completed a post-task survey. They were also given the option of a 5-minute break between sessions. Finally, we conducted a 20-minute semi-structured interview to explore their experience, their workflows, their perceived ownership of the revision, and their perspectives on AI assistance. The interview protocol is available in supplementary materials. Each study session lasted approximately 90 minutes. Participants received a $20 gift card as compensation for their time.

5.5 Measures

During the task sessions, we collected usage logs (i.e., participant actions with descriptions and timestamps) to obtain quantitative metrics on how users managed feedback and revised their essays. For the final versions, we recruited two experienced English teachers with extensive backgrounds in practicing, teaching, and assessing writing for an expert evaluation. They rated the degree of improvement in the revisions compared to the originals in a 7-point Likert scale, without awareness of the conditions under which the artifacts were produced. Following Choi et al.'s design [8], raters handled significant disagreements (>2 score difference) through discussion and re-evaluation ($ICC = 0.73$).

In the post surveys, we included ten questions based on our design goals (§3.1-3.3) to assess participants' perceived support from the systems. We also included one question to assess their perceived ownership in the AI-assisted writing experience. In addition, the survey included the NASA Task Load Index (NASA-TLX) [27] to assess the perceived effort required to use each system, along with five questions from the work of Wu et al. [74] to evaluate participants' self-perceived experience of using the AI system. All survey items used a 7-point Likert scale.

5.6 Analysis

To compare survey responses and ratings between conditions, we conducted statistical analysis using the Wilcoxon signed-rank test, given the ordinal nature of Likert-scale responses. For quantitative metrics of user behaviors and outcomes, we used the paired t-test. For qualitative analysis of interview transcripts, we followed established open-coding protocols [5, 60]. Two authors independently coded the transcripts, then discussed, reached a consensus, and created a consolidated codebook. This codebook was then used for thematic analysis to identify emerging topics from the interviews. The entire research team collectively reviewed the coding outcomes to refine high-level themes.

6 Findings

We began by analyzing the perceived effort required to use each system, measured by NASA-TLX scores ($M_S = 3.26$ $SD_S = 1.29$ *vs.* $M_B = 4.22$ $SD_B = 1.33$; $W = 18.500$ $p = 0.058$). The results

[4] The two essays were distributed to each participant one day in advance to allow them to familiarize themselves with the content before the study.

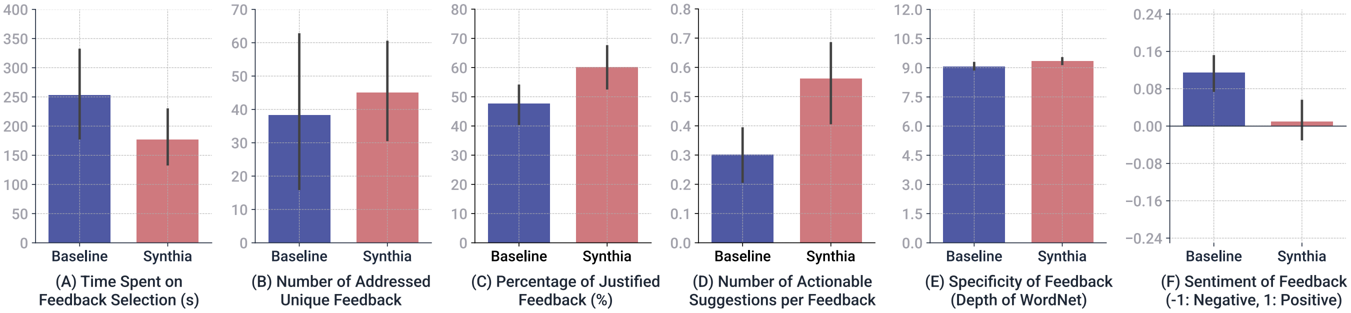


Figure 7: Bar plots illustrating the statistical metrics of participant performance of feedback interpretation in two conditions, where the t-values and p-values (*: p<.05, **: p<.01, ***: p<.001) from the Student's paired t-test are reported. Error bars represent 95% confidence intervals (CIs). Justified feedback (C) is defined as comments that users used to create revisions and that were rated justified using the method described in [43]. Similarly, actionable suggestions (D), specificity (E), and sentiment (F) were also identified and calculated following the same computational method proposed in [43].

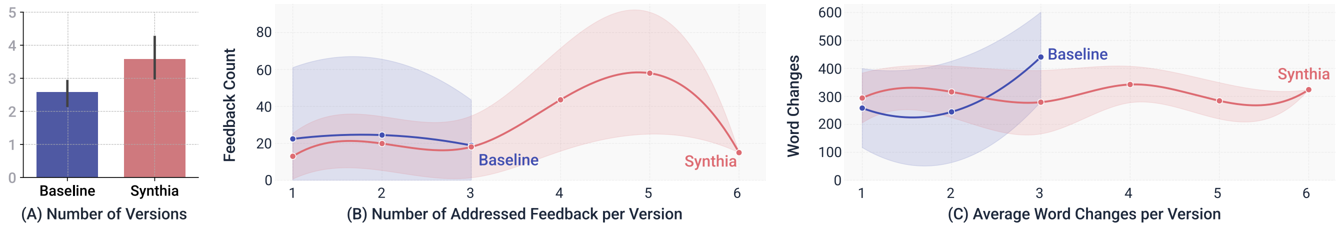


Figure 8: (A) Bar plots showing the number of versions created by participants across two conditions. (B) Line chart depicting how the number of addressed feedback evolves across versions in two conditions. (C) Line chart illustrating the average number of word changes per version in two conditions.

Table 1: The statistical metrics of participants' behavioral performance and outcomes of feedback-driven writing revision, where the t-values from the Student's paired t-test, W-values from the Wilcoxon signed-rank paired test (only for ownership), and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

Metrics		Synthia		Baseline		Statistics	
		M	SD	M	SD	$t(11)\,/\,W$	p
Feedback Interpretation	Time Spent on Feedback Selection (s)	177.29	89.47	253.59	138.61	-1.698	0.059
	# of Addressed Unique Feedback in All Versions	45.08	29.20	38.25	42.56	0.370	0.359
	% of Justified Feedback	60.16	13.68	47.71	11.68	2.418	**0.017***
	# of Actionable Suggestions per Feedback	0.56	0.26	0.30	0.17	2.476	**0.015***
	Specificity of Feedback (Depth of WordNet)	9.34	0.28	9.08	0.30	1.808	**0.049***
	Sentiment of Feedback (-1: Negative, 1: Positive)	0.01	0.08	0.11	0.07	-3.398	**0.003****
Revision Exploration	# of Explored Versions	3.58	1.08	2.58	0.67	2.345	**0.019***
	# of Word Changes in Final Versions	327.67	100.25	338.75	202.36	-0.188	0.573
	# of Addressed Feedback in Final Versions	23.00	24.37	9.67	11.99	1.952	**0.039***
	Expert Ratings of Final Versions (Content)	5.49	1.01	5.29	1.29	0.411	0.689
	Expert Ratings of Final Versions (Language)	5.79	0.87	5.17	1.50	1.321	0.213
	Perceived Ownership of Final Versions	5.33	1.72	4.42	1.93	10.000	**0.049***

showed participants felt significantly less rushed when working with Synthia (M_S = 2.92 SD_S = 1.73 *vs.* M_B = 4.42 SD_B = 1.38; W = 14.500 p = 0.029*) and experienced less frustration (M_S = 3.33 SD_S = 1.61 *vs.* M_B = 4.75 SD_B = 1.77; W = 13.000 p = 0.040*).

In the following sections, we will investigate our research questions in depth and present the corresponding findings.

6.1 RQ1: Interpreting & Prioritizing Feedback

As shown in Fig. 6, participants found Synthia significantly more helpful than Baseline in assisting feedback interpretation (Q1: $M_S = 5.25$ $SD_S = 1.55$ *vs.* $M_B = 3.42$ $SD_B = 1.62$; $W = 57.000$ $p = 0.018^{*}$). Specifically, they reported that the system made it easier to understand lengthy feedback (Q2: $M_S = 5.92$ $SD_S = 1.44$ *vs.* $M_B = 3.83$ $SD_B = 1.44$; $W = 59.000$ $p = 0.011^{*}$), categorize and organize feedback (Q3: $M_S = 6.17$ $SD_S = 1.19$ *vs.* $M_B = 3.08$ $SD_B = 1.62$; $W = 75.000$ $p = 0.003^{**}$), and prioritize valuable feedback to consider (Q4: $M_S = 5.00$ $SD_S = 1.71$ *vs.* $M_B = 2.92$ $SD_B = 2.02$; $W = 58.000$ $p = 0.014^{*}$). Moreover, Synthia improved participants' ability to understand the intentions behind feedback (Q5: $M_S = 5.25$ $SD_S = 1.66$ *vs.* $M_B = 3.17$ $SD_B = 1.53$; $W = 60.000$ $p = 0.009^{**}$) and identify similar perspectives across different responses (Q6: $M_S = 4.58$ $SD_S = 1.08$ *vs.* $M_B = 2.42$ $SD_B = 1.08$; $W = 45.000$ $p = 0.004^{**}$).

Beyond subjective ratings, behavioral data further illustrates the advantages of Synthia (Fig. 7). Participants spent comparable time selecting feedback when using Synthia ($M_S = 177.29s$ $SD_S = 89.47s$ *vs.* $M_B = 253.59s$ $SD_B = 138.61s$; $t(11) = -1.698$ $p = 0.059$), yet identified significantly higher-quality comments. Specifically, they recognized more justified feedback ($M_S = 60.16\%$ $SD_S = 13.68$ *vs.* $M_B = 47.71\%$ $SD_B = 11.68$; $t(11) = 2.418$ $p = 0.017^{*}$; Fig. 7C), more actionable feedback ($M_S = 0.56$ $SD_S = 0.26$ *vs.* $M_B = 0.30$ $SD_B = 0.17$; $t(11) = 2.476$ $p = 0.015^{*}$; Fig. 7D), and more detailed feedback ($M_S = 9.34$ $SD_S = 0.28$ *vs.* $M_B = 9.08$ $SD_B = 0.30$; $t(11) = 1.808$ $p = 0.049^{*}$; Fig. 7E). Interestingly, participants using Synthia were encouraged to address significantly less positive feedback ($M_S = 0.01$ $SD_S = 0.08$ *vs.* $M_B = 0.11$ $SD_B = 0.07$; $t(11) = -3.398$ $p = 0.003^{**}$; Fig. 7F). In contrast, when using Baseline, they tended to focus more on positive feedback, a trend aligned with prior research [55], suggesting that Synthia helps shift attention toward constructive criticism that may otherwise be overlooked.

We also find qualitative evidence of Synthia's support for feedback interpretation. For example, clustering and coloring by type helped participants categorize feedback and identify the most critical weaknesses in their writing (P4, P10, P12) without needing to "*read through every sentence to figure out what a feedback means.*" After coloring the bubbles by type, both P10 and P12 noticed that most of the bubbles were brown, which helped them identify a lack of evidence as a major issue and focus on addressing it. In addition, when describing their selection strategies, participants often prioritized less positive feedback in Synthia (P3, P8, P10), "*I think those flashy red bubbles are more urgent... I want to solve them first.*" (P8) This is in line with our observations of behavioral data. In contrast, participants using Baseline found it "*hard to decide whether a feedback is good or not*" (P10), leading them to act less strategically and ultimately select all feedback (P12).

6.2 RQ2: Exploring & Iterating on Revisions

As shown in Fig. 6, users found Synthia significantly more helpful than Baseline in identifying sections that needed revision (Q7: $M_S = 5.50$ $SD_S = 1.51$ *vs.* $M_B = 3.33$ $SD_B = 2.06$; $W = 60.000$ $p = 0.009^{**}$), rapidly testing different revision ideas (Q9: $M_S = 5.58$ $SD_S = 1.51$ *vs.* $M_B = 4.08$ $SD_B = 1.73$; $W = 53.000$ $p = 0.040^{*}$), and tracking and comparing their versions (Q10: $M_S = 5.92$ $SD_S = 1.31$ *vs.* $M_B = 4.5$ $SD_B = 2.15$; $W = 37.000$ $p = 0.048^{*}$).

6.2.1 Revision Behaviors. Participants using Synthia explored significantly more versions compared to Baseline (Fig. 8A; $M_S = 3.58$ $SD_S = 1.08$ *vs.* $M_B = 2.58$ $SD_B = 0.67$; $t(11) = 2.345$ $p = 0.019^{*}$). When examining the progression of versions more closely, we found that Synthia encouraged users to address more feedback over time (Fig. 8B), while maintaining a steady rate of word changes across versions (Fig. 8C). For each version with Synthia, participants iterated on the generated AI drafts, either through regeneration or manual edits, an average of 2.68 times ($SD = 2.18$).

In interviews, we asked participants to walk us through their revision process. They exhibited an evolving and experimental approach (P4, P9, P10, P12). For example, P12 initially focused on surface-level issues such as grammar but later refined their arguments by incorporating feedback related to reasoning and evidence. After examining the AI-generated drafts, P12 realized that some feedback on evidence introduced changes misaligned with their revision goals. As a result, in their final version, P12 adopted a more balanced approach: they decided to focus on fewer but more specific reasoning-type feedback, guided by the specificity metrics.

In contrast, participants using Baseline often experimented with full automation but ultimately retreated to more manual, selective revision strategies. While six participants (e.g., P2, P4, P7, P9, P10, P12) attempted to generate complete drafts using the AI assistant, only one (P2) ultimately submitted the automated version. Participants who abandoned this approach cited concerns during interviews about quality control, lack of transparency, and a sense of cognitive disconnect. Participants struggled to "*compare whether the (automated) revision is better or not*" (P10). P9 noted, "*There's no transparency in generation ... [after] a few seconds [it gives] the new paragraph. You just replace that with your original essay, which I don't really like.*" P12, who requested automated breakdowns of feedback and categorized changes, admitted, "*I kind of gave up the selection process,*" and expressed concern over losing ownership of their writing. Similarly, P4 found it difficult to prioritize or make sense of feedback, instead generating automated versions by arbitrarily selecting feedback from different reviewers. These findings further support the value of our system in scaffolding a deeper, more interpretable revision process while preserving user agency.

6.2.2 Revision Outcomes. We further analyzed the final versions submitted by participants. We observed no significant difference in the number of word changes made ($M_S = 327.67$ $SD_S = 100.25$ *vs.* $M_B = 338.75$ $SD_B = 202.36$; $t(11) = -0.19$ $p = 0.573$). However, participants using Synthia addressed significantly more feedback items in their final drafts ($M_S = 23.00$ $SD_S = 24.37$ *vs.* $M_B = 9.67$ $SD_B = 11.99$; $t(11) = 1.952$ $p = 0.038^{*}$). The expert ratings of final revisions produced with Synthia and the Baseline are comparable in both content ($M_S = 5.49$ $SD_S = 1.01$ *vs.* $M_B = 5.29$ $SD_B = 1.288$; $t(11) = 0.411$ $p = 0.689$) and language ($M_S = 5.79$ $SD_S = 0.87$ *vs.* $M_B = 5.17$ $SD_B = 1.50$; $t(11) = 1.321$ $p = 0.213$). This result is likely due to both systems relying on the same underlying language models to generate drafts, which also led participants to perceive similar levels of direct support for text revision (Fig. 6 Q8: $M_S = 5.17$ $SD_S = 1.75$ *vs.* $M_B = 5.17$ $SD_B = 1.47$; $W = 20.500$ $p = 0.784$). Notably, participants reported a significantly stronger sense of ownership over their revisions when using Synthia compared to Baseline ($M_S = 5.33\, SD_S = 1.72$ *vs.* $M_B = 4.42\, SD_B = 1.93; W = 10.000\ p = 0.049^*$). The significantly higher number of addressed comments and increased sense of ownership suggest that Synthia facilitated greater engagement and agency in the revision process.

6.3 RQ3: User Perceptions of System Support

When assessing their overall experiences collaborating with AI, participants rated that Synthia significantly helped them think through the kind of outputs they wanted to complete ($M_S = 6.00\ SD_S = 1.21$ *vs.* $M_B = 4.33\ SD_B = 1.83$; $W = 47.000\ p = 0.026^*$). Furthermore, they rated Synthia as significantly more transparent about how it arrived at its final results ($M_S = 5.50\ SD_S = 1.24$ *vs.* $M_B = 4.08\ SD_B = 1.68$; $W = 55.000\ p = 0.027^*$), which allowed them to better track its progress.

Participants highlight how Synthia engaged them in a transparent AI-assisted writing experience, emphasizing the importance of being able to verify and track information. When selecting feedback, Synthia created a "*text-to-text layer*" that mapped feedback with target sentences, allowing participants to predict the impact from selected feedback after AI revision (P8, P9, P12). Meanwhile, the connection between feedback and its original context also allowed participants to "*verify whether it matched with the predictions.*" (P8) Participants also liked how Synthia enabled them to track feedback and their selection strategies over time, as Synthia saved their customized visualization for each version (P9, P10). For example, P10 compared Synthia with Baseline, "*Synthia is more transparent because I know which feedback I choose. I can easily find my history of selected feedback, but in [Baseline], ...it was hard for me to relocate feedback.*"

6.4 RQ4: User Challenges and Feedback

Although the bubble visualization effectively aided feedback prioritization, participants expressed concerns that it "*became a dominant way to navigate feedback.*" (P8) The visualization may direct user attention to certain feedback bubbles, such as those with salient colors and sizes, like P9 described, "*Sometimes, I just click on the most outstanding or eye-catching bubbles.*" Moreover, despite a higher sense of transparency in Synthia, participants still felt it was a bit of a "*black box*" (P9) and wished to better understand the rationale behind the mapping between feedback, text, and generated edits. For instance, P12 suggested, "*When applying changes, I would appreciate it if Synthia can give me reasons why [AI] did it, so I can better assess whether human intervention is needed...*" Lastly, participants found it difficult to understand feedback without knowing the "*background*" (P3) of the feedback providers. Future experiments could offer more fine-grained visualization controls to convey skills and experience of feedback providers.

7 Discussion

7.1 Balancing Granular Control and Global Structure in Interactive Revision Tools

Synthia focuses on sentence-level revision, supporting bottom-up exploration of feedback. This design aligns with prior research that emphasizes the value of fine-grained revision [1, 32, 57, 70, 84], where writers can try out sentence edits, evaluate suggestions in context, and iteratively build toward improved drafts.

However, this bottom-up focus also presents certain limitations. Some feedback may call for broader rhetorical or structural changes, which often require top-down planning across paragraphs or even topics. While such global revisions can technically be achieved by modifying multiple sentences, a more flexible revision span may offer a smoother and more engaging authoring experience. This reflects a broader tension between granular flexibility and structural coherence. Similar design considerations arise in exploratory data analytic tools [61] and malleable interfaces [52], where users benefit from switching between detail-focused and overview-oriented perspectives to maintain both precision and contextual understanding. Thus, we see design opportunities for future systems to bridge this gap. For instance, integrating outline-level overviews (such as the visual programming-style interface in VISAR [85]) and enabling users to fluidly shift between local and global perspectives (e.g., tracking how revisions affect narrative arcs [9]) could help writers reason about how sentence-level edits align with broader revision intentions.

7.2 Preserving Cognitive Reflection in AI-Augmented Revision

Recently, the HCI community continues to debate the rapid transformation of creative work driven by LLMs [12], which can generate text at unprecedented speeds [69]. This ease of generation brings a potential trade-off: while automation boosts efficiency, it may reduce opportunities for learners to independently develop revision skills. Writing expertise is often cultivated through slow, reflective, and effortful practice [13], which is a process that generative tools might inadvertently short-circuit.

Although our study did not directly measure reflection, participants reported that Synthia helped them better think through the kinds of revisions they wanted to make (§6.3), and their comments indicated strategic deliberation when revising with Synthia (§6.2). By structuring and visualizing feedback, Synthia may reduce the cognitive burden of sensemaking, potentially freeing up mental resources for deeper reflection on writing goals and tradeoffs.

We emphasize, however, that reflection is a potential outcome, not a guaranteed effect, of AI-assisted revision. Our tool augments feedback sensemaking, but deeper reflection may require complementary tools or scaffolds. For instance, once Synthia helps users identify high-value feedback worth further investment, tools like Friction [82] can support deeper, structured reflection on individual comments. Future HCI research should explore this automation–reflection tradeoff more systematically, examining how different forms of AI support either scaffold or suppress critical thinking during the revision process.

7.3 Visual Scaffolding for Exploratory Feedback-driven Revision

Synthia treats feedback not as static content to be consumed but as material for exploration. Its visual scaffolding supported participants in identifying relevant critiques, interpreting intent, and experimenting with revision strategies. Below, we reflect on how our design supports scalable sensemaking and opens new directions for feedback-centered tools.

From Rigid to Customizable Feedback Visual Encodings: Prior tools have used visual indicators to highlight feedback properties such as topics, sources, or sentiment [16, 79]. However, these signals are often fixed or predefined. In contrast, Synthia introduces customizable visual encodings (i.e., bubbles) that allow writers to dynamically reconfigure how feedback is grouped, highlighted, and prioritized based on evolving revision goals. Much like exploratory data analysis, this enables sensemaking at scale: participants actively shaped how they viewed the feedback to assess helpfulness, surface similarities, and compare competing suggestions.

Supporting Feedback Uniqueness and Complementarity: Although our system was designed for feedback recipients, we envision that visualizing feedback metrics can also scaffold feedback providers [58]. For example, by dynamically visualizing in-progress feedback, similar to live word count or spelling indicators, the visualization in Synthia could provide formative signals that nudge providers toward more actionable and constructive comments. In addition, the visual layout (e.g., clustering and arcs) could help reviewers notice overlapping critiques or detect under-addressed areas. If a cluster is already dense, for instance, contributors might be encouraged to offer novel or complementary insights, thereby improving the diversity and coverage of feedback across a group.

Toward Traceable, Negotiable Feedback Loops: Synthia's bidirectional linking across feedback, source text, and revisions surfaces relationships that are often hidden in linear revision tools. This traceability enables users to track how specific critiques are addressed (or not) during the revision process. It also opens opportunities for meta-feedback: writers could comment on individual feedback bubbles, flag helpful suggestions, or respond with clarifications, enabling lightweight, asynchronous dialogue. While past systems have aggregated critiques by region or theme [51, 76], Synthia adds a temporal and revisional dimension: showing not just what was critiqued, but how that critique influenced writing changes over time. Such functionality could foster mutual understanding, reduce redundancy, and promote deeper engagement within feedback-centric platforms.

7.4 Limitations

Our system has a few limitations. First, since we designed it around argumentative writing, its feedback categorization may not generalize to other genres like fiction, where critiques usually focus on characters, settings, and story lines. However, our workflows, visualizations, and scaffolds are applicable across different writing domains. Second, while our current design supports feedback clustering by provider ID, it does not yet distinguish between source types or sociocultural backgrounds, which can influence how feedback is interpreted and acted upon. Future iterations should incorporate richer metadata and visualization features to better surface such contextual cues during revision.

Our study methodology also has limitations. We followed previous work [79] to provide participants with existing essays and feedback to control revision difficulty. However, they might respond differently if they were revising their own work based on feedback they had received personally. In the future, we will investigate how personal investment influences system usage by having participants revise their own essays. Additionally, our user study only involves 12 participants. A larger and more diverse participant pool would provide a stronger foundation for evaluating the system's usability and generalizability. Finally, while we observed short-term benefits, a longitudinal study is needed to assess long-term writing skill development.

8 Conclusion

In conclusion, this paper presents Synthia, a system that visually scaffolds feedback-based writing revision with LLM-powered synthesis. Our study demonstrates that Synthia effectively enhances feedback-based revision by making critique more interpretable and actionable while fostering an intentional and transparent experience. By visually organizing feedback into interactive clusters and enabling bidirectional linking, Synthia empowers writers to curate, explore, and integrate feedback more strategically. The ability to generate alternative drafts further supports experimentation and iterative refinement. Overall, we believe Synthia offers promising implications for future revision tools that scaffold not just rewriting, but the interpretation, navigation, and application of feedback in all its complexity.

Acknowledgments

We sincerely thank our participants for sharing their thoughts and suggestions on our system, and we are grateful to all reviewers for their valuable insights and feedback.

References

[1] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445683

[2] Tazin Afrin and Diane Litman. 2018. Annotation and Classification of Sentence-Level Revision Improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 240–246. doi:10.18653/v1/W18-0528

[3] Bryan Anthony Bardine and Anthony Fulton. 2008. Analyzing the Benefits of Revision Memos during the Writing and Revision Process. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 81, 4 (March 2008), 149–154. doi:10.3200/TCHS.81.4.149-154

[4] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. arXiv:2309.10433 [cs] doi:10.1145/3613904.3642406

[5] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[6] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445372

[7] Ruijia Cheng, Ziwen Zeng, Maysnow Liu, and Steven Dow. 2020. Critique Me: Exploring How Creators Publicly Request Feedback in an Online Critique Community. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020), 161:1–161:24. doi:10.1145/3415232

[8] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2023. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. arXiv:2312.11949 [cs]

[9] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human*

Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3491102.3501819

[10] Amy Cook, Jessica Hammer, Salma Elsayed-Ali, and Steven Dow. 2019. How Guiding Questions Facilitate Feedback Exchange in Project-Based Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300368

[11] Patrick Crain, Jaewook Lee, Yu-Chun Yen, Joy Kim, Alyssa Aiello, and Brian Bailey. 2023. Visualizing Topics and Opinions Helps Students Interpret Large Collections of Peer Feedback for Creative Projects. *ACM Trans. Comput.-Hum. Interact.* 30, 3 (June 2023), 49:1–49:30. doi:10.1145/3571817

[12] Michele Cremaschi, Max Dorfmann, and Antonella De Angeli. 2024. A Steampunk Critique of Machine Learning Acceleration. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 246–257. doi:10.1145/3643834.3660688

[13] Arthur Cropley. 2006. In Praise of Convergent Thinking. *Creativity Research Journal* 18, 3 (July 2006), 391–404. doi:10.1207/s15326934crj1803_13

[14] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel P. Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-Writing with Language Models. arXiv:2402.11723 [cs]

[15] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3573–3590. arXiv:2203.03802 [cs] doi:10.18653/v1/2022.acl-long.250

[16] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion Space: A Scalable Tool for Browsing Online Comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1175–1184. doi:10.1145/1753326.1753502

[17] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me?: Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 229–239. doi:10.1145/3301275.3302265

[18] Linda Flower, John R. Hayes, Linda Carey, Karen Schriver, and James Stratman. 1986. Detection, Diagnosis, and the Strategies of Revision. *College Composition and Communication* 37, 1 (1986), 16–55. jstor:357381 doi:10.2307/357381

[19] Eureka Foong, Darren Gergle, and Elizabeth M. Gerber. 2017. Novice and Expert Sensemaking of Crowdsourced Design Feedback. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 45:1–45:18. doi:10.1145/3134680

[20] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:1803.07640 [cs]

[21] 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 (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. doi:10.1145/3532106.3533533

[22] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642139

[23] Gabriela Goldschmidt. 2011. Avoiding Design Fixation: Transformation and Abstraction in Mapping from Source to Target. *Journal of Creative Behavior* 45, 2 (June 2011), 92–100. doi:10.1002/j.2162-6057.2011.tb01088.x

[24] Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. ACM, Glasgow United Kingdom, 235–244. doi:10.1145/2757226.2757249

[25] Qingyu Guo, Chao Zhang, Hanfang Lyu, Zhenhui Peng, and Xiaojuan Ma. 2023. What Makes Creators Engage with Online Critiques? Understanding the Role of Artifacts' Creation Stage, Characteristics of Community Comments, and Their Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581054

[26] Maralee Harrell. 2005. Grading According to a Rubric. *Teaching Philosophy* 28, 1 (2005), 3–15. doi:10.5840/teachphil200528111

[27] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, Amsterdam, The Netherlands, 139–183. doi:10.1016/S0166-4115(08)62386-9

[28] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (March 2007), 81–112. doi:10.3102/003465430298487

[29] John R. Hayes. 2004. What Triggers Revision? In *Revision Cognitive and Instructional Processes*, Gert Rijlaarsdam, Linda Allal, Lucile Chanquoy, and Pierre Largy (Eds.). Vol. 13. Springer Netherlands, Dordrecht, 9–20. doi:10.1007/978-94-007-1048-1_2

[30] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3641895

[31] Yi-Ching Huang, Hao-Chuan Wang, and Jane Yung-jen Hsu. 2018. Feedback Orchestration: Structuring Feedback for Facilitating Reflection and Revision in Writing. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 257–260. doi:10.1145/3272973.3274069

[32] Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. 2020. Langsmith: An Interactive Academic Text Revision System. arXiv:2010.04332 [cs]

[33] Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the Rough: Generating Fluent Sentences from Early-Stage Drafts for Academic Writing Assistance. arXiv:1910.09180 [cs]

[34] Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jack Jamieson, and Kentaro Inui. 2023. Use of an AI-Powered Rewriting Support Software in Context with Other Tools: A Study of Non-Native English Speakers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3586183.3606810

[35] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3581196

[36] David G. Jansson and Steven M. Smith. 1991. Design Fixation. *Design studies* 12, 1 (1991), 3–11.

[37] Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the Human Revision Process in Scientific Writing. arXiv:2210.15067 [cs]

[38] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3586183.3606737

[39] Hyeonsu B. Kang, Gabriel Amoako, Neil Sengupta, and Steven P. Dow. 2018. Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. doi:10.1145/3173574.3174180

[40] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. 2017. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 233–245. doi:10.1145/2998181.2998196

[41] Taewook Kim, Hyomin Han, Eytan Adar, Matthew Kay, and John Joon Young Chung. 2024. Authors' Values and Attitudes Towards AI-Bridged Scalable Personalization of Creative Language Arts. arXiv:2403.00439 [cs] doi:10.1145/3613904.3642529

[42] Stephen King. 2000. *On Writing: A Memoir of the Craft.* Simon and Schuster, New York, NY, USA.

[43] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M. Gerber, Brian P. Bailey, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4627–4639. doi:10.1145/3025453.3025883

[44] Saeed Latifi, Omid Noroozi, Javad Hatami, and Harm J.A. Biemans. 2021. How Does Online Peer Feedback Improve Argumentative Essay Writing and Learning? *Innovations in Education and Teaching International* 58, 2 (March 2021), 195–206. doi:10.1080/14703297.2019.1687005

[45] Saeed Latifi, Omid Noroozi, and Ebrahim Talaee. 2021. Peer Feedback or Peer Feedforward? Enhancing Students' Argumentative Peer Learning Processes and Outcomes. *Brit J Educational Tech* 52, 2 (March 2021), 768–784. doi:10.1111/bjet.13054

[46] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L. C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Agnia Sergeyuk, Antonette Shibani, Disha Shrivastava, Lila Shroff, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia H. Rho, Shannon Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. arXiv:2403.14117 [cs] doi:10.1145/3613904.3642697

[47] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. doi:10.1145/3491102.3502030

[48] Yoonjoo Lee, Tae Soo Kim, Minsuk Chang, and Juho Kim. 2022. Interactive Children's Story Rewriting Through Parent-Children Interaction. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 62–71. doi:10.18653/v1/2022.in2writing-1.9

[49] Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 553–562.

[50] Astute Digital Solutions Ltd. 2025. Ref-n-Write. https://www.ref-n-write.com/.

[51] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 473–485. doi:10.1145/2675133.2675283

[52] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable Overview-Detail Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–25. doi:10.1145/3706598.3714164

[53] Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Towards Automated Document Revision: Grammatical Error Correction, Fluency Edits, and Beyond. arXiv:2205.11484 [cs]

[54] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. doi:10.1145/3544548.3581023

[55] Thi Thao Duyen T. Nguyen, Thomas Garncarz, Felicia Ng, Laura A. Dabbish, and Steven P. Dow. 2017. Fruitful Feedback: Positive Affective Language and Source Anonymity Improve Critique Reception and Work Outcomes. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1024–1034. doi:10.1145/2998181.2998319

[56] Omid Noroozi, Harm Biemans, and Martin Mulder. 2016. Relations between Scripted Online Peer Feedback Processes and Quality of Written Argumentative Essay. *The Internet and Higher Education* 31 (2016), 20–31.

[57] Mohi Reza, Nathan Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2023. ABScribe: Rapid Exploration of Multiple Writing Variations in Human-AI Co-Writing Tasks Using Large Language Models. arXiv:2310.00117 [cs] doi:10.48550/arXiv.2310.00117

[58] Jeffrey Rzeszotarski and Aniket Kittur. 2012. CrowdScape: Interactively Visualizing User Behavior and Output. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. Association for Computing Machinery, New York, NY, USA, 55–62. doi:10.1145/2380116.2380125

[59] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Creativity and Cognition*. ACM, Virtual Event Italy, 1–10. doi:10.1145/3450741.3465253

[60] Raymond Scupin. 1997. The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization* 56, 2 (1997), 233–237. jstor:44126786

[61] Robert Spence and Lisa Tweedie. 1998. The Attribute Explorer: Information Synthesis via Exploration. *Interacting with Computers* 11, 2 (Dec. 1998), 137–146. doi:10.1016/S0953-5438(98)00022-8

[62] Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Junichi Tsujii and Jan Hajic (Eds.). Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1501–1510.

[63] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3613904.3642400

[64] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3586183.3606756

[65] Yu-Chia Tseng and Chao Zhang. 2025. The Role of Politeness Strategies in Online Design Feedback Exchange. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3706599.3720100

[66] Selen Türkay, Daniel Seaton, and Andrew M. Ang. 2018. Itero: A Revision History Analytics Tool for Exploring Writing Behavior and Reflection. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3170427.3188474

[67] Peter D. Turney. 2006. Similarity of Semantic Relations. *Computational Linguistics* 32, 3 (2006), 379–416.

[68] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and Prevention of Large Language Model Use in Crowd Work. arXiv:2310.15683 doi:10.48550/arXiv.2310.15683

[69] Florent Vinchon, Todd Lubart, Sabrina Bartolotta, Valentin Gironnay, Marion Botella, Samira Bourgeois-Bougrine, Jean-Marie Burkhardt, Nathalie Bonnardel, Giovanni Emanuele Corazza, Vlad Glăveanu, Michael Hanchett Hanson, Zorana Ivcevic, Maciej Karwowski, James C. Kaufman, Takeshi Okada, Roni Reiter-Palmon, and Andrea Gaggioli. 2023. Artificial Intelligence & Creativity: A Manifesto for Collaboration. *The Journal of Creative Behavior* 57, 4 (2023), 472–484. doi:10.1002/jocb.597

[70] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376732

[71] Bruce Weigl. 1976. Revision as a Creative Process. *English Journal* 65, 6 (Sept. 1976), 67–68. doi:10.58680/ej197614763

[72] Naomi E. Winstone, Robert A. Nash, Michael Parker, and James Rowntree. 2017. Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. *Educational Psychologist* 52, 1 (Jan. 2017), 17–37. doi:10.1080/00461520.2016.1207538

[73] Naomi E. Winstone, Robert A. Nash, James Rowntree, and Michael Parker. 2017. 'It'd Be Useful, but I Wouldn't Use It': Barriers to University Students' Feedback Seeking and Recipience. *Studies in Higher Education* 42, 11 (Nov. 2017), 2026–2041. doi:10.1080/03075079.2015.1130032

[74] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–22. doi:10.1145/3491102.3517582

[75] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A Visual Interactive System to Enhance the Persuasiveness of Arguments in Online Discussion. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 319:1–319:30. doi:10.1145/3555210

[76] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1433–1444. doi:10.1145/2531602.2531604

[77] Yu-Chun Grace Yen, Steven P. Dow, Elizabeth Gerber, and Brian P. Bailey. 2016. Social Network, Web Forum, or Task Market?: Comparing Different Crowd Genres for Design Feedback Exchange. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, Brisbane QLD Australia, 773–784. doi:10.1145/2901790.2901820

[78] Yu-Chun Grace Yen, Steven P. Dow, Elizabeth Gerber, and Brian P. Bailey. 2017. Listen to Others, Listen to Yourself: Combining Feedback Review and Reflection to Improve Iterative Design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. ACM, Singapore Singapore, 158–170. doi:10.1145/3059454.3059468

[79] Yu-Chun Grace Yen, Joy O. Kim, and Brian P. Bailey. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376380

[80] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. doi:10.1145/3490099.3511105

[81] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1017. doi:10.1145/2818048.2819953

[82] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–27. doi:10.1145/3706598.3714316

[83] Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A Corpus of Annotated Revisions for Studying Argumentative Writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1568–1578. doi:10.18653/v1/P17-1144

[84] Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A Web-Based Revision Assistant for Argumentative Writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, John DeNero, Mark Finlayson, and Sravana Reddy (Eds.). Association for Computational Linguistics, San Diego, California, 37–41. doi:10.18653/v1/N16-3008

[85] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–30. doi:10.1145/3586183.3606800