Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection

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

Kexin Ju
Cornell University
Ithaca, New York, USA
kj234@cornell.edu

Peter Bidoshi
Cornell University
Ithaca, New York, USA
pjb294@cornell.edu

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

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

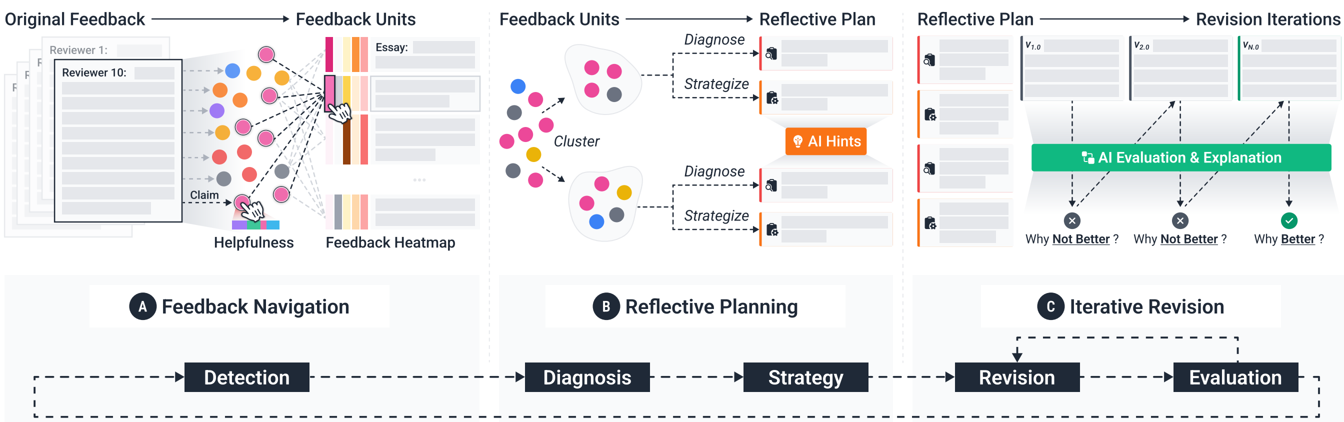

Figure 1: Design Framework of Friction. Users engage in a reflective cycle through three key stages to revise their essays based on feedback. In Feedback Navigation (A), Friction breaks down and categorizes feedback, highlighting potential areas for revision through a co-located heatmap, aiding in the detection of writing issues. In Reflective Planning (B), Friction encourages translating feedback into actions by providing adaptive AI hints for diagnosing issues and strategizing improvements. In Iterative Revision (C), Friction offers AI-driven evaluation and suggestions, encouraging continuous reflection and iterations.

Abstract

This paper introduces Friction, a novel interface designed to scaffold novice writers in reflective feedback-driven revisions. Effective revision requires mindful reflection upon feedback, but the scale and variability of feedback can make it challenging for novice writers to decipher it into actionable, meaningful changes. Friction leverages large language models to break down large feedback collections into manageable units, visualizes their distribution across sentences and issues through a co-located heatmap, and guides users through structured reflection and revision with adaptive hints and real-time evaluation. Our user study ($N = 16$) showed that Friction helped users allocate more time to reflective planning, attend to more critical issues, develop more actionable and satisfactory revision plans, iterate more frequently, and ultimately produce higher-quality revisions, compared to the baseline system. These findings highlight the potential of human-AI collaboration to foster a balanced approach between maximum efficiency and deliberate reflection, supporting the development of creative mastery.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Feedback, Reflection, Sensemaking, Writing, Revision, Creativity, Large Language Models

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714316

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3706598.3714316

1 Introduction

Revision is essential for high-quality writing. It leads writers to reassess, reorganize, and reconceptualize their work [42]. Effective revision typically involves gathering feedback from stakeholders, reflecting on the identified issues, and making changes to address them [6, 74, 113]. Nowadays, writers can obtain feedback not only from instructors and peers, but also from online communities [49] and crowdsourcing platforms [120] quickly and affordably. These platforms improve availability of feedback and greatly expand its scale to dozens or hundreds of individual pieces of free-form text responses. This rapid access to diverse feedback offers transformative potential in creative writing, especially for novices who have historically struggled to obtain varied critiques for revision.

However, easy access to feedback can be a double-edged sword. Even with only a dozen feedback responses, each one may contain critiques addressing different aspects of the writing, creating potentially hundreds of more issues for authors to address individually. For example, when an argumentative essay is critiqued, a reviewer might point out the need for stronger evidence to support key claims, suggest clarifying the thesis statement, and recommend addressing potential counterarguments—all within the same textual comment. With additional readers, the volume of critiques can expand exponentially, covering elements such as claims, warrants, evidence, and rebuttals. The diversity of opinions can also increase, with the same element receiving different suggestions from different reviewers. This can leave novice writers overwhelmed, unsure where to begin, and struggling to translate feedback into improvements [42, 59].

To execute effective revisions, writers must first organize feedback comments based on common issues, identify problematic areas from a group of similar critiques, and then engage in a reflective planning process by diagnosing specific deficiencies and formulating a revision strategy [42]. Once this plan is in place, revision often involves multiple cycles, as authors continuously refine and reassess their work until the issues are considered resolved [32]. However, novices often lack the necessary skills and experience to navigate this complex, iterative process [26, 42, 122].

Current research in Human-Computer Interaction (HCI) has primarily focused on generating feedback summaries or suggested changes by AI for rapid revision [1, 62, 80, 95, 107]. For example, students can input the feedback they received to ChatGPT and ask it to regenerate the full article based on the feedback. While these approaches have shown promise in helping writers revise efficiently, several challenges remain: 1) focusing on high-level summaries can obscure potentially valuable individual comments and may collapse the distribution of issues in a writing sample, making it harder to identify problematic areas for broader improvement; and 2) introducing easy alternatives short-circuits the reflective learning process and poses the risk of deskilling, leaving fewer opportunities for novices to practice their writing skills and bring in their own perspectives.

We aim to *strike a balance between efficiency and reflection*, minimizing the strain of dealing with the complexity of feedback while exposing writers to the benefits of reflective practice. We believe that writers could benefit from a more deliberate, slowed-down pace of interaction, which encourages deeper reflection. Prior work has shown that such reflection leads to higher quality results [121]. However, most learners do not naturally engage in meaningful reflection without targeted intervention [8, 17, 21, 23, 83]. The challenge lies in achieving a desirable difficulty level [9] of reflection by scaffolding writers.

Generative AI offers unique potential here: it can alleviate information overload by revealing patterns among critiques and coach users through the process of reflection by offering adaptive hints. Based on the writing context and feedback, these hints can pinpoint specific issues and suggest actionable solutions. Rather than replacing writers in the revision process, generative AI can quickly evaluate revised content and encourage iterative improvements. By augmenting sensemaking for initial information processing steps and then slowing down parts of the process that require reflection, this approach is likely to foster deliberate practice in revision behaviors, potentially delivering more sustainable learning outcomes for writers in the long run [4, 35, 71].

In this paper, we introduce FRICTION[1](Figure 1)—a novel Generative AI-powered interface designed to scaffold novice writers in structured, in-depth reflection on feedback for critical revisions. First, it breaks down large feedback collections into manageable units, identifies writing problems, and predicts their distribution across the essay. This information is presented via a co-located heatmap (Figure 1A) to help users navigate feedback and structurally address a certain type of issue across different sections. It also visualizes helpfulness metrics to help users prioritize feedback. Upon selecting a problematic sentence, FRICTION prompts users to cluster similar feedback units and plan actions for each cluster, maintaining a detailed focus on valuable comments. If users struggle with diagnosing problems or proposing solutions, FRICTION offers adaptive hints for inspiration (Figure 1B). As users revise their sentences, FRICTION evaluates the improvements and provides explanations to encourage continuous reflection and iteration (Figure 1C). In addition to outlining the design and development of FRICTION, we will show that our generative model pipelines were able to provide useful support with low risk of error or miscommunication through a technical evaluation. While this paper focuses on argumentative writing as an example domain for system design, our approach is applicable across various writing genres that require feedback.

To evaluate FRICTION, we conducted a within-subjects study ($N = 16$) and compared FRICTION with a baseline consisting of a chat assistant and a similar interface without key features of our system. Results indicated that FRICTION enabled participants to navigate feedback more quickly, encouraging them to dedicate more time to reflective planning. The feedback heatmap was reported as particularly effective in enhancing participant sensemaking of both global and local issues in their writing. During reflective planning, participants addressed nearly three times as many feedback instances,

[1] FRICTION is an acronym for **F**eedback to **R**evision with A**I** Support in A**ction**. The name was inspired by the educational philosophy of introducing desirable difficulties in learning that require considerable, but desirable, effort to improve long-term achievement [9].

with over 80% targeting deep, content-related issues, and produced longer, more actionable, and more satisfactory action plans. When revising sentences, they iterated more times per sentence, which resulted in higher quality revisions and greater satisfaction with the outcomes. Our findings suggest that human-AI interactive systems can strike a balance between accelerating tasks and encouraging deliberate reflection, helping users to develop true creative mastery.

In summary, this paper presents the following contributions:

- a new approach that leverages AI to support writers in reflecting on and acting upon received feedback during the revision process, rather than replacing writers by regenerating the entire text;
- Friction[2], an AI-powered writing support system that implements this approach by helping novice writers organize, navigate, and act on feedback through features like an interactive feedback heatmap, AI-generated action hints, and real-time iteration evaluation;
- empirical findings from a within-subject user study that demonstrate the effectiveness of Friction in fostering reflective behaviors and facilitating revision.

2 Related Work

2.1 Feedback Tools

In creative fields like design, writing, and music, gathering feedback from diverse audiences is essential for achieving successful outcomes. To streamline these feedback cycles, HCI researchers have developed a range of techniques for efficiently collecting high-quality feedback at scale [7, 22, 48, 70, 124]. For example, crowdsourcing platforms like Voyant [118] and CrowdCrit [84] enable creators to obtain structured feedback from crowds. However, people from diverse backgrounds and areas of expertise may prioritize issues differently. Even when evaluating the same design elements, they can offer varying, sometimes contradictory, opinions [55].

The complexity of feedback makes its effectiveness dependent on recipients' ability to interpret, learn, and act on it. Prior research has shown that novices are less likely than experts to recognize key insights from feedback for improvement [43]. Instead, they often rely on personal preference, focusing on feedback that is positive, easy to implement, or aligning with their existing ideas. Thus, despite having access to high-quality feedback, they tend to make superficial changes rather than addressing deeper content issues [47, 82, 85].

To support feedback interpretation, previous work has introduced methods that aggregate feedback into issue categories to help users navigate large-scale feedback [59, 84, 118]. For example, ReviseO [59] categorizes feedback as related to semantics, language, or mechanics. However, writing feedback goes beyond responding to different issues and often hints at areas that need to be revised. Identifying problematic areas and assessing their severity are often more difficult for novices [42], especially when feedback is given in free-form text from online communities or crowdsourcing platforms. To address it, our work visually aggregates feedback along two dimensions: issue type and the targeted areas of the text, enabling users to both understand feedback holistically and apply it more effectively to specific sections of their writing.

[2] We open-sourced our code at https://github.com/zhangchaodesign/friction to promote reproducibility and community engagement.

Another line of research has invented new visual representations for the semantic characteristics of textual feedback [25, 118, 122]. For instance, Decipher [25, 122] visualizes the topic and sentiment structure of feedback, aiding novice designers in interpreting feedback from multiple sources. While these visual approaches are effective in helping users understand the sources and content of feedback, they often stop short of guiding users through the process of planning and implementing their revisions.

In contrast, our work introduces a framework that supports novice writers by providing scaffolding for feedback interpretation, reflective planning, and iterative revision. Our system walks users through the process of translating feedback into actionable revisions, helping them move beyond merely understanding feedback to effectively applying it in practice.

2.2 Reflection Interventions

Creativity can be learned through reflection. Schön [99] introduced the concept of "reflection-in-action," emphasizing the importance of regular reflection for professional creators to think critically as new information, such as feedback, becomes available. Similarly, Handley et al. [52], in their work *Beyond 'doing time'*, underscore the importance of deliberate reflection in the process of engaging with feedback. They argue that simply collecting and skimming through feedback is insufficient; true value arises when feedback is mindfully reflected upon, interpreted, and used to deepen understanding, ultimately leading to changes in behavior.

Reflection occurs throughout the revision process [41]. It induces cognitive dissonance, reconciling writers' current output with their intended text [31, 40, 65] and thereby promoting effective action on feedback and the development of writing skills [63, 97, 100]. However, most learners do not deliberately engage in meaningful reflection without intervention [8, 17, 23, 83, 128], suggesting that a reflection process needs to be explicitly guided within the system design.

Researchers, therefore, designed various interventions for learners, such as reflection prompts with hints [17] and a toolbox of reflective strategies [39]. For example, Chi et al. [15] found that prompting students to self-explain each step of the reflective process causes higher learning gains than having them study the material without such prompting. Jackson et al. [63] noted improvements in students' feedback utilization and academic performance when required to formulate action plans based on their coursework feedback. Choi et al. [17] demonstrated the effectiveness of reflective hints to increase delayed knowledge transfer and learner perception of learning. However, these pedagogical materials often fail to provide personalized guidance tailored to each student's unique writing challenges and diverse feedback.

With the rapid advances of large language models (LLMs), HCI researchers have explored a few new ways of leveraging its generative capacity to support reflection. For example, Jamplate [119] integrates LLM-generated responses into design template, helping designers reflect on their creative ideas. These prior systems provided initial evidence of the effectiveness of LLMs in supporting reflection, which inspired us to explore its adaptability in feedback-driven revision scenarios.

2.3 Intelligent Writing Interfaces

The HCI community has a long-standing interest in designing interactive writing interfaces [78]. These interfaces support writers in various writing stages, ranging from brainstorming ideas [45, 98, 125–127] and planning outlines [132] to drafting content [18, 30, 58, 64, 73, 123] and refining existing text [1, 62, 80, 95, 107].

Among writing revision tools, commercial applications like Microsoft Word and Grammarly[3] generate suggestions for users to address convention or grammar, fluency, and organization changes. Ref-N-Write[4] enhances this by focusing on stylistic improvements for more professional or academic writing. Academically, tools designed to aid novices in scientific writing suggest more fluent sentence alternatives [32, 60–62, 67, 87] and typically present users with multiple revision options [11, 60, 79, 95, 123]. Each option is often accompanied by a confidence score to help users select the most appropriate variant [36, 44]. While these approaches are effective for quick revision, they offer fewer opportunities for authors, especially novices, to practice their writing skills in the presence of easy shortcuts. Importantly, writers may lose ownership with AI taking more agency in generating content [30]. Inspired by Zhou and Sterman's concept of *Creative Struggle* [133] in AI-assisted writing, our work aims to balance enabling creative momentum by offloading efforts in feedback sensemaking, while still allowing enough creative struggle for writers to develop skills.

Instead of suggesting changes, some tools generate automatic visual or numerical assessments for writing quality [88, 117]. For example, AL [111] visualizes the relations of sentences in an argument and scores their persuasiveness,. ArgRewrite [130] tracks and assesses changes made by a writer for each sentence. Sentences represent a natural boundary of text, which allows for clearer demarcations where edits are needed, facilitating easier tracking and application of changes. Following previous sentence-based revision interfaces [60, 111, 130], we also set sentences as the scope of revision spans. In addition, Sterman et al. [105] developed a visual interface that allows users to explore and analyze styles across large bodies of literature. Their work showed that presenting visualization alongside writing fosters self-dialogue and critical reflection. Inspired by this idea, we aggregate and visualize large collections of feedback in a co-located heatmap to help writers actively interpret feedback and identify areas for improvement.

In conclusion, previous research has primarily focused on suggesting alternatives, auto-complete features, or automatic assessment, without considering external human feedback in writing revision. To address this gap, our work develops an AI-infused interface aimed at supporting the common practice of revising based on collected textual feedback. Rather than suggesting easy alternatives, our tool guides users in creating justified and actionable revision plans and evaluating their revisions to encourage iterative improvements, preserving opportunities for reflection and skill development.

3 The Friction System

In this section, we will first draw from past research to frame the workflow of feedback-driven revision and introduce the design goals that Friction seeks to achieve. Then, we will demonstrate the main features of Friction, which support users at each leverage point in the revision workflow—namely, detection, diagnosis, strategy, revision, and evaluation—by illustrating how a novice writer, Lee, would interact with the system to revise her essay. Lastly, we elaborate on the implementation of Friction and an evaluation to examine the capabilities of its technical pipelines.

In our demonstration of Friction, we chose argumentative writing as an example domain to help design the system. Argumentative writing is a task that requires individuals to present and defend their perspective on a specific issue or topic. It is particularly important in educational settings, where it is commonly assigned to novice writers to help them practice and develop essential writing skills [135]. To effectively convey their argument, writers must engage in clear, structured reasoning, consider other viewpoints, and transform their stance into a persuasive written piece. This process offers a valuable opportunity for feedback and improvement, enabling learners to strengthen their claims, refine supporting evidence, and enhance their overall writing ability [76, 77, 89].

3.1 Feedback-Driven Revision Workflows: Learning from Expert Practices

To design an effective system, we drew from past research on the specific workflow that experts follow when conducting feedback-driven revision [42, 59, 72, 122]. Pushing students to adopt expert practices has been shown to lead to positive educational outcomes, because they learn new and effective strategies for improving their work [29, 92]. Following this thread of research, we extrapolate the step-by-step improvements observed from expert practices: *detection*, *diagnosis*, *strategy*, *revision*, and *evaluation* (Figure 2). This five-stage framework forms the foundation of our tool's user flow.

Experts begin by *detecting* problematic areas in the text. At this initial stage, authors organize the received feedback and connect it with their writing [97]. They break down feedback into manageable comments and then filter, cluster, and prioritize them [122]. They also map comments to sections in their writing that potentially need revisions. Each comment can be associated with multiple sections, and each section can be linked to multiple comments. Next, experts *diagnose* the specific issues within a particular section highlighted by the feedback [42]. This includes distilling specific issues in the targeted area from general comments, determining the root causes, and establishing achievable goals for revision. Following diagnosis, they devise *strategies* to address diagnosed issues [42]. This involves selecting appropriate actions based on their understanding of the problem, how the intended text should be formed, and the feedback received. They then proceed to the *revision* stage, where they iteratively revise the problematic section following their planned actions. Finally, they *evaluate* the effectiveness of their revisions based on the goals they set [91]. This evaluation promotes continuous reflection, encouraging writers to continually refine their writing [100].

3.2 Design Goals

While experts navigate the stages of revision with ease, it remains challenging and counterintuitive for novices [42]. To inform our system design for non-experts, we gained insights from the literature into (1) the challenges faced by novices in feedback

[3]https://www.grammarly.com/
[4]https://www.ref-n-write.com/

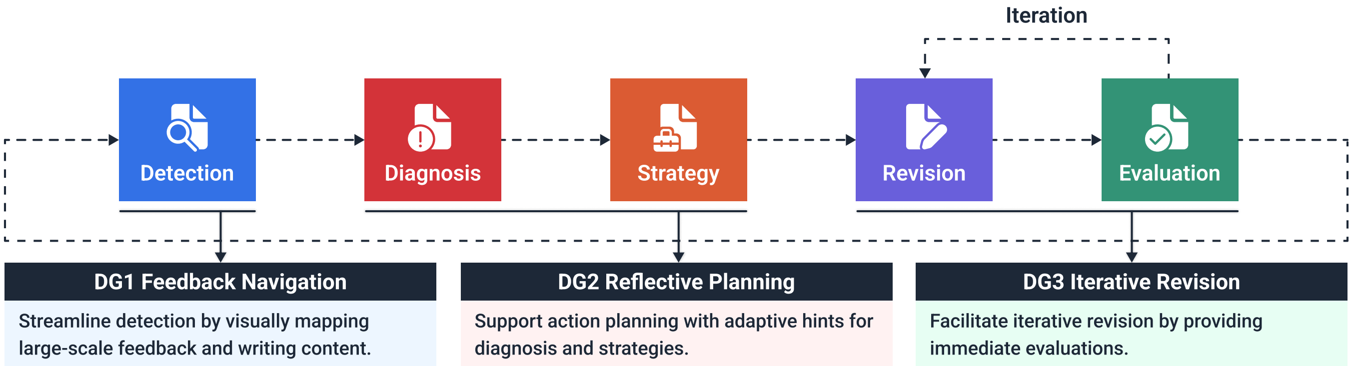

Figure 2: System workflow and design goals of Friction: In the first step (i.e., Detection), our system aims to provide a visual structure that maps feedback and writing, simplifying the organization of feedback and the detection of problematic content. In the second and third steps (i.e., Diagnosis and Strategy), our system aims to guide novices in revision planning by offering hints that pinpoint specific issues and suggest actionable solutions. In the fourth and fifth steps (i.e., Revision and Evaluation), our system aims to provide immediate evaluations after each iteration to inform users' decisions on further revisions.

sensemaking, creative reflection, and writing revision, and (2) the capabilities and limitations of existing feedback and writing tools. As such, we identified the following design goals (DG) for different stages of our system workflow:

3.2.1 DG1: Streamline detection by visually mapping large-scale feedback and writing content. Different types of feedback target various writing skills and text sections, creating a high cognitive demand for interpretation, especially with large amounts of feedback [93]. Novices face the dual challenge of organizing large-scale feedback and linking feedback to specific text needing revision. Research shows novices are less skilled than experts at identifying improvement opportunities in their work and understanding their writing deficiencies when reviewing feedback [43]. They often struggle to grasp "what feedback is for" [114, 115]. Previous research suggested that integrating feedback into revisions usually requires adding structure to the content [118, 122]. Thus, DG1 is about to provide a visual structure that maps and aggregates large-scale feedback and writing, simplifying feedback organization and the detection of problematic content.

3.2.2 DG2: Support action planning with adaptive hints for diagnosis and strategies. Novices often struggle to translate feedback into actions [114], leading to frustration and "behavioral disengagement" [52]. This difficulty stems from a lack of knowledge about strategies and opportunities for implementing feedback [114]. Jonsson's review [68] highlights this knowledge gap as a key reason for students' poor use of feedback. Students need immediate, applicable direction rather than just invitations to use support [94]. Providing actionable hints has proven effective in education [17, 65]. The hints should be targeted and adaptable to the writing content and the writing issues. Thus, DG2 suggests the system design to guide novices in effective action planning by providing actionable, adaptive hints for diagnosis and strategies.

3.2.3 DG3: Facilitate iterative revision by providing immediate evaluations. Revision is an iterative process where writers test different alternatives to improve their text. Evaluation is key to revealing dissonance between the current output and the intended text [31, 40, 65], motivating reflection and enabling effective iterations. However, novices often struggle to evaluate their own work, making it difficult to understand the impact of their changes [42]. Typically, writers seek external evaluations to guide their revisions [26]. Effective evaluations should identify improvements (e.g., better or not) and analyze the effectiveness of their implementation based on the goals they set [91]. Thus, DG3 sets a goal for our system to provide immediate *evaluations* after each iteration to inform users' decisions on further *revisions*.

3.3 User Scenario

To demonstrate our system, we present an imagined scenario in which Lee, an English as a Second Language (ESL) student aiming to improve her writing skills, uses Friction to revise an essay she recently practiced for an upcoming IELTS test[5]. She is tasked with arguing "*whether we should place less emphasis on technological solutions and more on other values.*" After drafting, she sends her essay to her English teacher and three classmates, and posts it on EssayForum.com[6] for feedback. She received ten reviews in total, each containing one or two paragraphs of critique. While Lee is pleased with the extensive and constructive feedback, she finds it difficult to revise her work due to the overwhelming and varied nature of the comments. Therefore, she decides to upload her essay and the collected reviews to Friction and use it to aid her revision.

Friction consists of an Essay Panel (Figure 3A), a Feedback Heatmap (Figure 3B), and a Reflection & Revision Panel (Figure 3C). Lee will go through three different pages in the Reflection & Revision Panel (Figure 3C.1) as she navigates through the feedback, makes reflective plans, and revises the content, respectively. In the following sections, we will first introduce the features of Friction

[5]IELTS, or International English Language Testing System, is a well-recognized English language proficiency test for non-native speakers. In its Writing Task 2, test takers write an argumentative essay. We have shortened the prompt from a sample test for use in this scenario.
[6]EssayForum.com is a non-profit online community for ESL students to solicit feedback on their essays.

at each stage followed by an example scenario depicting how Lee will use FRICTION to decipher feedback into improvements step by step (Figure 4).

3.3.1 Feedback Navigation. First, FRICTION breaks lengthy reviews into smaller, individual feedback units and displays them as cards in the Reflection & Revision Panel. A feedback unit is defined as one or more sentences that describe a coherent thought [122]. Users can check the original feedback of each unit by hovering over "Original Feedback" (Figure 3C.3).

Each feedback unit is classified into one of the eight categories of writing issues proposed by Zhang et al. [129]: surface issues (i.e., *Conventions*, *Word-usage*, and *Organization*) and content issues (i.e., ***C**laim*, ***W**arrant*, ***E**vidence*, ***R**ebuttal*, and ***G**eneral content*). Prioritization is a particularly challenging aspect of action planning [82, 85], yet it is closely linked to writing improvement [131]. Without support, novice writers tend to focus on making superficial changes, rather than addressing deeper content issues [84]. Following prior work [1, 130], we encourage writers to focus more on content issues by color-coding the borders of feedback cards based on category: cool colors (e.g., blue) for surface-level feedback and warm colors (e.g., orange) for content-related feedback (Figure 3C.5).

Feedback Overview. For each feedback unit, FRICTION predicts the sentences that need revision. With this set of predictions, FRICTION provides a feedback overview (Figure 3C.2) to reveal the three most serious categories of writing issues. For each of the three categories, it displays the number of feedback units and the number of targeted sentences. This overview helps novices quickly form initial impressions about how reviewers perceived their writing deficiencies.

Feedback Distribution. FRICTION surfaces distribution patterns of feedback. All feedback units are aggregated in a co-located heatmap (Figure 3B), which visually organizes the unit by different sentences into five content-related categories of writing issues (**DG1**). This helps novices effectively *detect* areas that require attention and systematically address a certain type of issue across different sections. The horizontal axis represents issue categories, while the vertical axis indicates sentence locations. The color depth of each cell indicates the number of units for that specific category and sentence. The heatmap, the Essay Panel, and the Reflection & Revision Panel are cross-linked. For instance, users can click on a sentence to view all comments related to it. They can also click on a specific cell in the heatmap (Figure 4 a) to see comments associated with a particular category and the sentence in that row. Solving a feedback unit in the Reflection & Revision Panel will also result in a color depth reduction in the heatmap.

Feedback Prioritization. To help users prioritize helpful feedback, FRICTION evaluates each feedback unit and displays a horizontally stacked histogram at the top-left corner of each feedback card (Figure 3), indicating the unit's overall helpfulness. This visualization incorporates four helpfulness metrics proposed by Krause et al. [13, 75] (Figure 3C.4)—*Negativity* (i.e., the degree of negative feedback), *Actionability* (i.e., the number of actionable suggestions), *Justification* (i.e., the extent of explanations provided), and *Specificity* (i.e., the level of detail). Users can hover over each cell to view a textual description and the normalized score (ranging from 0 to 1) for each metric. In addition, FRICTION provides a toolbar for users to search, sort, or filter feedback units by keywords, sources, category, and helpfulness metrics. Users can also remove disagreeable units by clicking the dismiss icon.

Example Scenario. In our design scenario (Figure 4), Lee begins by checking the feedback overview. She sees that "*Evidence*" is one of the three most common issues mentioned by reviewers and then is intrigued by a particularly dark cell in the heatmap located in the seventh sentence row, indicating that this sentence faces the most significant evidence-related issues according to the color mapping (Figure 4 a). Clicking on this cell, FRICTION displays all relevant feedback units, which Lee then sorted by "*Specificity*" using the toolbar. She hovers over the stacked histogram bars of each unit to carefully read the detailed descriptions of helpfulness. Dismissing two units with a low specificity score, Lee decides to focus on the remaining high-quality feedback to develop plans for revising the sentence.

3.3.2 Reflective Planning. By clicking the "Go to Plan" button (Figure 3C.6), users will proceed to the second phase in the Reflection & Revision Panel (Figure 3C.1), which involves developing revision plans to address feedback units concerning a selected sentence. The planning process is structured into three steps: *cluster*, *diagnose*, and *strategy*, during which FRICTION offers tailored AI support to help users self-explain their understanding of addressing feedback and develop justified and actionable plans for each feedback cluster (**DG2**). Figure 4B illustrates an example of the interface at this stage of the Reflection & Revision Panel, prior to moving to the next step.

Clustering Feedback. Before developing revision plans, experts typically organize feedback into meaningful groups and form a high-level view of opinions present in the feedback set [122]. Thus, FRICTION first prompts users to organize feedback cards into clusters by dragging and dropping feedback units into different groups. Once clustered, the feedback is collapsed to reduce visual clutter, with detailed information accessible via hovering (Figure 4 b). For each cluster, FRICTION provides a summary (Figure 4 c), which highlights broader patterns and connections among feedback units, making it easier for users to address related issues comprehensively.

Diagnosing Problems. Novices often struggle to identify the core problems of a specific sentence that the feedback points to, especially when feedback is general in nature. Without a clear understanding of the problems, effective revisions are difficult to implement. Thus, FRICTION prompts users to "*diagnose the specific problem within this sentence*" for each cluster in a text box. If users find it difficult to pinpoint specific issues, they can request AI hints by clicking a button (e.g., Figure 4 d). These hints are tailored to the essay's context and the feedback provided. They highlight potential problems the user may have overlooked, helping them arrive at a more specific and accurate diagnosis of the underlying issues.

Devising Strategies. The final stage in the process is for users to devise solutions that address the diagnosed problems. Novices, however, often lack knowledge of appropriate strategies for effectively implementing feedback [114]. To address this, FRICTION prompts users to "*devise a solution that can address the diagnosed problem*" for each cluster in a text box and provides AI-generated hints to assist in formulating strategies specific to the sentence and its associated

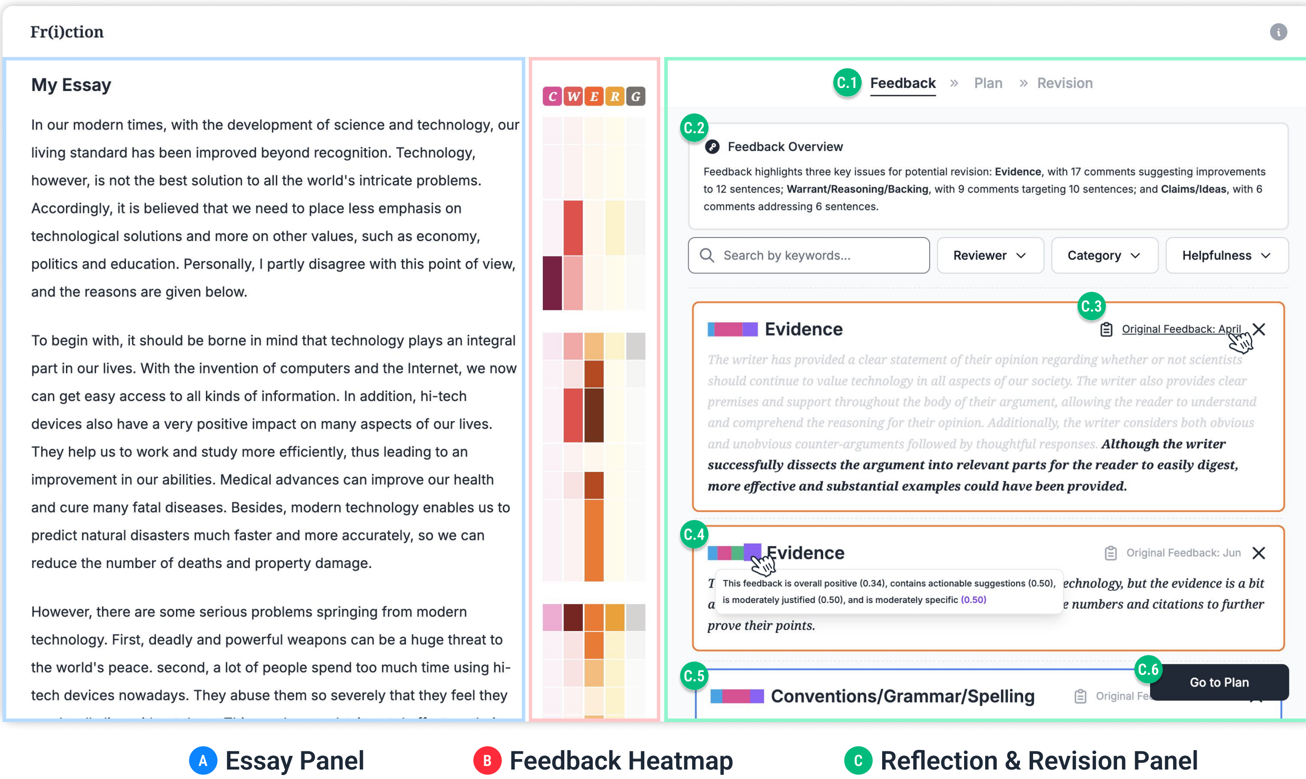

Figure 3: User Interface of Friction. (A) The Essay Panel displays the user's uploaded essay. (B) The Feedback Heatmap visualizes the distribution of feedback across individual sentences and five content-related writing issue categories. (C) In the Reflection & Revision Panel, users can review feedback, develop reflective plans, and make revisions to their work.

issues (e.g., Figure 4 ⓔ). These adaptive hints offer revision techniques, suggest examples, or provide guidance on rephrasing or restructuring problematic sentences. By shuffling the hints with a button click, users are exposed to various approaches, empowering them to develop practical, contextually appropriate solutions.

Example Scenario. In our design scenario (Figure 4), Lee begins the reflective planning process by grouping the second, fourth, fifth, and eighth comments together (Figure 4 ⓑ). Friction immediately provides a summary highlighting common themes in this cluster, such as the need for concrete connections between evidence and argument (Figure 4 ⓒ). Next, Lee starts to diagnose the problem. She reads the cluster carefully but still finds the feedback too general. By clicking the hint button (Figure 4 ⓓ), she receives suggestions such as "*This sentence uses general adjectives like 'very' and 'many' but doesn't specify the extent of the positive impact or which specific aspects are affected*," which help her realize that this sentence is too obscure and lacks specific evidence. With a clear diagnosis in mind, Lee moves on to propose a solution to address the problem, but still unsure of how to rephrase the sentence with more precise details and supporting evidence. She clicks the hints button for strategy ideas (Figure 4 ⓔ). Friction suggests "*providing academic statistics to demonstrate how hi-tech devices positively impact our lives in various areas, such as education, health, or communication*," which inspires Lee to include statistics she recently learned from a study about health to revise the sentence.

3.3.3 Iterative Revision. With action plans, Friction encourages users to iteratively *revise* the sentence. It *evaluates* the improvements between the revised and the original sentences. If the revised sentence shows improvement, the sidebar of the revision panel turns green (Figure 4 ⓖ) and Friction will provide explanations on how the solutions are successfully implemented in the revision; otherwise, it remains gray while Friction will explain why there are no improvements and suggest next steps. The iterative revision history is recorded for users to review. This iterative evaluation by Friction will help novices reflect on their improvements, understand the impact of their changes, and make effective iterations (**DG3**).

Example Scenario. In our design scenario (Figure 4), Lee revises the sentence: "*In addition, hi-tech devices also have a very positive impact on many aspects of our lives*" based on the plan she made in the last phase. However, her first attempt, "*Also, high-tech devices are really good for healthcare, like because of a 30% increase in recovery...*," is flagged as inadequate by the gray sidebar. Undeterred, Lee carefully reviews the analysis from Friction (Figure 4 ⓕ), which states, "*...Instead of presenting a clear, credible statistic, the revision uses informal language and vague phrases, diminishing the academic tone...*" She then makes several other attempts, closely

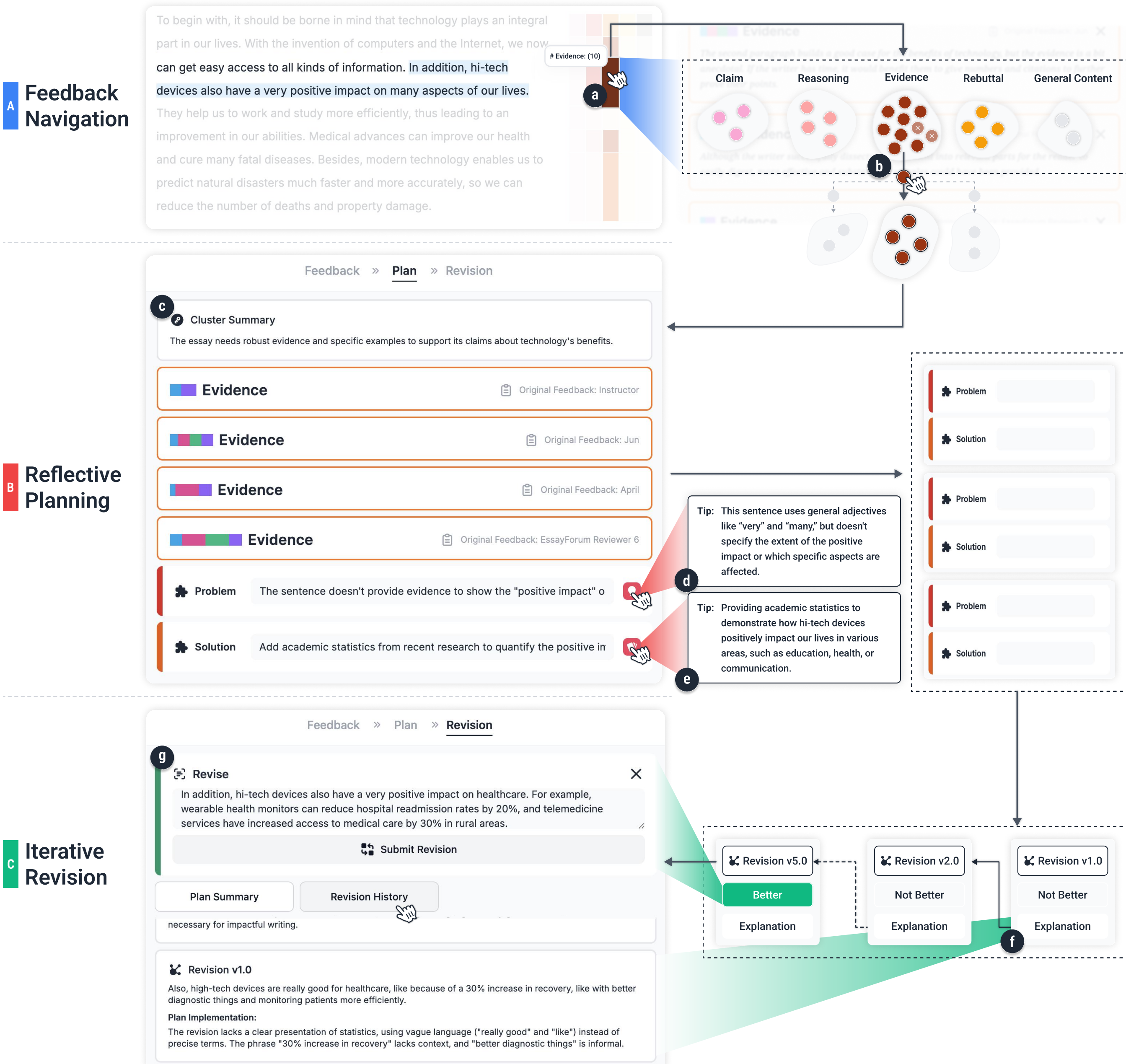

Figure 4: Design scenario of FRICTION. (A) Co-located heatmap of the second paragraph of Lee's essay during Feedback Navigation: FRICTION categorizes and breaks down feedback, using a heatmap to highlight areas needing revision. Lee's attention is drawn to the darkest cell, located at the intersection of the seventh sentence row and the evidence column. (B) Summary, feedback, and action plan for one feedback cluster in Reflective Planning: FRICTION guides users to decipher feedback into actionable plans with adaptive AI hints. Lee diagnoses issues and strategizes improvements based on these hints. (C) Revision history and final successful revision of Lee in Iterative Revision: FRICTION provides real-time AI evaluation of each revision, motivating Lee to continually refine her work.

following the evaluation by Friction, and adjusting accordingly. Finally, the sidebar turns green (Figure 4 g), and Lee writes a successful revision: "*hi-tech devices have a very positive impact on healthcare. For example, wearable health monitors can reduce hospital readmission rates by 20%, and...*" Through this iterative process, Lee gains a deeper understanding of her writing flaws and learns how to apply targeted evidence.

Later, with the help of Friction, Lee revises other problematic sentences, reflecting deeply on the feedback, which leads to a gradual and sustainable improvement in her overall writing skills.

3.4 Implementation

The web app of Friction is implemented in React.js. It integrates Firebase for storing log data, Python Flask for the back-end server, and OpenAI API to access a large language model (GPT-4).

We prompt GPT-4 to break down feedback, classify feedback purposes, predict problematic sentences (§3.3.1), generate hints for diagnosis and strategies (§3.3.2), and analyze plan implementation (§3.3.3). We followed the work of Wu et al. [116] to break down complex tasks into a set of LLM primitive operations. Each prompt consists of three parts: directing the model to adopt the role of an "*English writing teacher*" to set the tone; using the default prompting structures proposed by Wu et al. [116] to give explicit instructions; and injecting up-to-date information on the essay, feedback, and plans monitored from user actions as context. For special tasks like hint generation, we list a set of descriptive criteria of good reflection derived from previous work [17, 29, 81] to guide the model. Sample prompts and outputs are available in Appendix A.

To assess the improvements in each iteration (§3.3.3), we adapted the Revision Quality dataset by Afrin and Litman [2] to fine-tune GPT-3.5 via the OpenAI fine-tune API[7]. The dataset comprises 940 pairs of original and revised sentences, each labeled to indicate whether the revision is "better" or "not better" than the original sentence. We divided the dataset into training (60%), validation (20%), and test sets (20%). The training and validation sets were used to fine-tune our model. The test set was used to evaluate the model performance, as detailed in §3.5.3.

3.5 Technical Evaluation

We validate the efficacy of our LLM pipelines, acknowledging that they may be prone to hallucinations or other inaccuracies [66], which may backfire by misguiding users or diminishing the overall usability of the system. Specifically, we evaluate Friction's ability to (1) accurately classify feedback units, (2) detect problematic sentences, diagnose writing issues, and suggest revision strategies, and (3) assess the improvements of revised sentences.

The feedback dataset was made through crowdsourcing, a method proven to be an effective source of high-quality feedback, comparable to that from social media, online communities, and even experts [120, 121, 124]. To ensure the quality of the data, we adhered to best practices by selecting workers from Prolific[8], with a 99% approval rating and a minimum of 1,000 previous submissions. Participants were limited to native English speakers in the United States with at least a Bachelor's degree. We sampled 100 feedback units on 6 essays from crowd workers as the dataset to evaluate Friction. The process of feedback collection and some discussion of its potential limitations are detailed in Appendix B.1.

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

3.5.1 Performance of LLM Pipelines in Feedback Purpose Classification. Two research assistants annotated the purposes of the samples in the dataset, achieving an initial Cohen's Kappa of 0.80. After resolving 13 conflicts, the ground-truth consists of 12% *Organization*, 3% *Word-usage*, 14% *Conventions*, 15% *Claim*, 20% *Warrant*, 26% *Evidence*, 7% *Rebuttal*, and 3% *General content/Others* feedback. After prediction, our prompted model achieved an overall precision of 0.90, recall of 0.84, and a macro F1-score of 0.84. A closer analysis showed that although several unique or rare units in the *General content/Others* category (33% accuracy) were misclassified, the model achieved good accuracy in all other categories: 100% in *Word-usage*, *Evidence*, and *Rebuttal*, 93% in *Conventions*, 87% in *Claim*, 83% in *Organization*, and 75% in *Warrant*.

3.5.2 Performance of LLM Pipelines in Detection, Diagnosis, and Strategies. As there is no established ground truth for performance evaluation of reflective planning, we conducted an expert rating to assess the quality of generation from our LLM pipeline. Given all feedback samples and the corresponding essays, the LLM pipeline produced an average of 2.6 problematic sentence-issue pairs per sample and generated 3 revision strategies to address each pair. Three evaluators rated the quality of generation for each feedback on a 5-point Likert scale, achieving an inter-rater reliability of 0.84. The results demonstrated that our tool performed well in detecting problematic sentences (M = 4.41, SD = 0.60), diagnosing writing issues (M = 4.64, SD = 0.53), and devising revision strategies (M = 4.35, SD = 0.62). Evaluators provided detailed comments, highlighting units where our method excelled or underperformed to explain their scores. Detailed comments and cases are displayed in Appendix B.2.

3.5.3 Performance of LLM Pipelines in Evaluating Revision Improvements. We used the test set of Revision Quality dataset [2] in §3.4 to evaluate the performance of the fine-tuned GPT-3.5 and compare it to a prompted GPT-4 model as the baseline. The fine-tuned model outperformed the prompted GPT-4, achieving high levels of precision (*ours*: 0.89 vs. *baseline*: 0.88), recall (*ours*: 0.86 vs. *baseline*: 0.75), and a macro F1-score (*ours*: 0.87 vs. *baseline*: 0.81). These results demonstrated the effectiveness of our fine-tuned model in accurately evaluating the improvements of sentence-level revisions.

4 Evaluation

To further evaluate the efficacy of Friction, we conducted a within-subjects controlled experiment[9] which compares fully-featured Friction against a baseline version of Friction with 16 novice writers. We focus on novices in our initial evaluation as

[7] https://platform.openai.com/docs/guides/fine-tuning
[8] https://www.prolific.com/
[9] The study received approval from our institution's IRB.

suggested by previous research [42, 122]: this audience currently has the most to gain from tools for feedback-driven revision. In other words, novices are most likely to use a system like FRICTION in the real world and most likely to show benefits in an experimental setting. For this study we will examine immediate behavioral changes that FRICTION evokes among these participants as an initial proof of efficacy. In the future we hope to explore learning effects more deeply to understand how FRICTION can support skill development over long periods of use.

The baseline system shared a similar interface (Appendix C.3) with FRICTION but without its major advancements, such as feedback heatmap, AI hints, and AI evaluation, etc. In this baseline system, users could manually select and cluster feedback pieces, create action plans, and make necessary revisions on certain sentences. AI hints were replaced with validated reflection prompts from the work of Yen et al. [121] to resemble established practices for supporting novices in reflection. Based on the prompts, participants can write their revision plans for each feedback cluster in a text box in a similar manner to FRICTION. In addition, there is the possibility that general AI functionality could be responsible for any improvements offered by FRICTION over control rather than the specific affordances that FRICTION contributes, so we provided participants in baseline with access to the GPT-4 version of ChatGPT[10] to level out this difference. ChatGPT, a generative AI chatbot based on the same model as FRICTION, is now commonly used for writing assistance [106]. Users can upload text files to ChatGPT and provide instructions for processing the text.

4.1 Hypothesis

Drawing on Flower et al.'s framework [42], our study anticipates that FRICTION will bridge the knowledge and intention gaps of novices, thereby improving their performance in both reflection and revision. For example, with a feedback heatmap that will help them quickly organize feedback and locate problematic sentences, we expect that they will reflect on more feedback and revise more sentences. Additionally, by offering AI hints, FRICTION ought to lower the barriers for novices to tackle content-level feedback and improve the quality of crafted reflective plans [37, 38]. That is to say, their diagnoses are expected to be more justified, and their strategies are expected to be more actionable [81]. Moreover, AI evaluation of revised sentences is anticipated to encourage novices to reconcile their current efforts with their intended text outcomes, thus facilitating revision iterations [31, 40]. High-quality planning and multiple iterations should further improve the quality of revision. Therefore, we make the following hypotheses with regard to participants' performance:

- ***H1*** Compared to the baseline, FRICTION will significantly enhance novices' performance in reflective planning (***H1***). This improvement will be evidenced by a higher number of feedback units being addressed (*H1a*), an increased proportion of content-level feedback among those units (*H1b*), and reflective plans that are longer (*H1c*), more justified (*H1d*), more actionable (*H1e*), and more satisfactory (*H1f*).
- ***H2*** Compared to the baseline, FRICTION will significantly enhance novices' performance of iterative revision (***H2***). This improvement will be demonstrated by an increased number of revised sentences (*H2a*) and word changes (*H2b*), more iterations per revised sentence (*H2c*), higher expert ratings for revision quality (*H2d*), and higher satisfaction levels with their revised sentences (*H2e*).

[10]https://chat.openai.com/

4.2 Methodology

4.2.1 Participants. We recruited 16 participants (8 female, 8 male) aged 22—35, with one preferring not to say their age (M = 24.53, SD = 3.31), from our university via email advertising and word-of-mouth. None of the participants had advanced knowledge of the project. All participants self-reported as non-native English speakers and ESL writers. We specifically recruited ESL writers, following the approach of Huang et al. [59], who also recruited ESL writers as novices for similar tasks. Subjectively, the participants rated their writing expertise on a scale from 1 (beginner writer) to 7 (professional writer), yielding an average self-rated expertise of 3.25 (SD = 0.68). Based on the two criteria, we anticipated that most, if not all, of our participants would be novice writers. Additionally, all participants reported that they often used generative AI tools, especially ChatGPT, in their daily writing practice. Each participant received a 20-dollar gift card as compensation for their time. Appendix C.1 provides the detailed information of our participants.

4.2.2 Task Materials. Participants received two essays written by ESL learners from a widely used dataset [104]. Both essays were comparable in scope—one about technology development and the other about mobile phones—and each was approximately 300 words long. Following the method in Appendix B.1, we collected 10 pieces of feedback for each essay via crowdsourcing, totaling around 2,000 words per set. The three trained expert evaluators from technical evaluation rated the feedback's perceived usefulness on a 7-point Likert scale (ICC = 0.77). An independent t-test showed no statistical difference, indicating no observable quality difference between the two feedback sets (5.2 *vs.* 5.3; $t(18) = 0.22, p = .826$)

4.2.3 Study Procedure. The user study procedure is outlined in Figure 5. In the beginning, the researchers collected informed consent and demographic information from the participants. Following this, participants engaged in two separate sessions, each starting with a 3-5 minute tutorial and followed by a 20-minute task session using either FRICTION or baseline. We used a within-subjects Latin square experimental design [96] to counterbalance the order of materials and system conditions. At the beginning of each session, the tutorials highlighted the key features of the current tool, and the following task required participants to review the feedback, reflect, and make revisions by using the tool. As participants all had prior writing experience with ChatGPT, we avoided mentioning any specific cases or prompts for how to use ChatGPT to avoid influencing their natural approach and to ensure unbiased results. Sessions were designed to be independent, with no carryover effects between them. After each session, they completed a post-task survey. Participants also had the option of a 10-minute break between sessions. Lastly, we conducted a 20-minute semi-structured interview to ask about the difference between their experience in the two conditions, their workflows, their

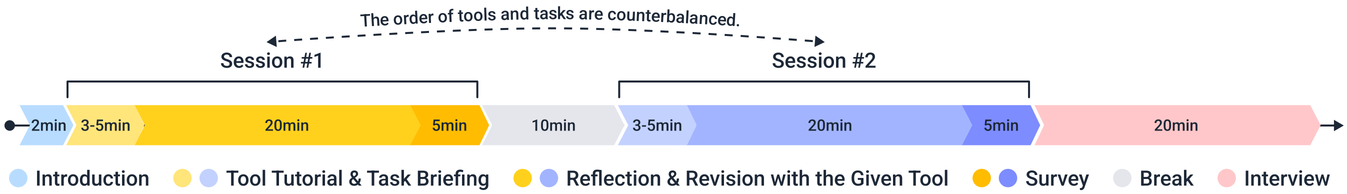

Figure 5: Study Procedure.

perceived ownership towards final revision, and their perception of the tools. The interview protocol is presented in Appendix C.2.

4.2.4 Measures. We gathered usage logs (i.e., participant actions with descriptions and timestamps) to obtain quantitative metrics for user behaviors. We used this data to calculate the time allocated for each stage and observed how often participants checked AI hints and how they used these hints in reflective planning. We also collected their reflective plans and revised sentences. With these materials, we calculated the number of feedback units addressed (both content-level and surface-level), the number of sentences revised, and the number of iterations, among other metrics.

The post-session survey included questions regarding the overall usefulness of the given system, as well as specific inquiries into the utility of Friction's inner functions (e.g., feedback heatmap, AI hints on diagnosis/strategy, AI evaluation on iterations). Additionally, the survey gauged participants' satisfaction with their reflective plans and sentence revisions. Lastly, the survey incorporated several standardized scales for a comprehensive evaluation: the short-form User Engagement Scale (UES) [90] and the Task Load Index (NASA-TLX) [54]. Because revision is traditionally viewed as a creative process [12, 112] and to explore whether users can effectively express their own ideas with the involvement of AI, we also included the Creativity Support Index (CSI) [14]. All items in the post-session survey adopt a 7-point Likert scale.

We also conducted an expert evaluation of the revisions. Two experts, who were both English teachers with extensive years of experience in practicing, teaching, and assessing writing, were recruited for this task. They were presented with 96 pairs of original and revised sentences (3 randomly selected pairs × 16 participants × 2 conditions), along with the associated feedback. The task is to rate the degree to which the revisions addressed the feedback on a 5-point Likert scale, without awareness of the conditions under which the artifacts were produced. Following Choi et al.'s design [16], raters handled significant disagreements (>2 score difference) through discussion and re-evaluation.

5 Results

In our quantitative analysis, we explore how well participants did in each condition and the performance characteristics of Friction. We analyzed participants' usage logs, reflection and revision outcomes, and data collected from surveys using statistical methods. In qualitative analysis of interview transcripts, we followed established open-coding protocols [10, 101]. Two authors first independently coded the transcripts. After that, they discussed, reached a consensus, and created a consolidated codebook. This codebook was then used to conduct a thematic analysis to identify emerging topics from the interviews. The entire research team collectively reviewed the coding outcomes to discern high-level themes.

5.1 Reflective Planning Performance

We first calculated the number of feedback units being addressed by participants in each condition. Feedback was considered resolved when participants created a revision plan for it and the final revision showed improvements as measured by our revision improvement evaluation model (§3.5.3) We expect that more effective tool will lead to a higher number of units addressed (*H1a*). As shown in Figure 6, participants attended to about 6.6 more feedback units when using Friction compared to using the baseline. Using a paired t-test, we observed that there is a statistically significant difference ($t(15) = -4.54, p < .001^{***}$), supporting our hypothesis. Next, we calculated the percentage of content-level feedback units among the reflected feedback units. Prior work suggests that novices tend to address surface-level issues [47, 65, 85], so we investigated whether Friction helped participants to examine deeper, content-level features. We applied the same model used to classify feedback in Friction to classify the planned feedback in baseline (see §3.4). Our findings revealed that in the Friction condition, over 80% (9 out of 10.38 per participant) of feedback addressed by participants pertained to essay content, which is significantly higher than in the baseline condition ($t(15) = -5.31, p < .001^{***}$; *H1b*). These findings support our hypotheses and indicate that Friction facilitated effective reflective planning to address feedback, especially for content-level issues.

We then analyzed the quality of reflective plans written by participants. Here we use length as a proxy measure for the general efforts that participants have put into the process, as supported by the tools. As shown in Figure 6, the reflective plans in Friction were approximately twice as long as those in baseline. A paired t-test showed that this was significant for both diagnosis ($t(15) = -2.92, p = .011^{*}$) and strategies ($t(15) = -3.57, p = .003^{**}$), supporting *H1c*. Then, we calculated the average number of sentences in participants' reflective plans that contained justified explanations (denoted as justified diagnosis) and the average number of sentences that included actionable solutions (denoted as actionable strategies) [13, 75, 81]. Participants developed significantly more actionable strategies when using Friction compared to the baseline ($t(15) = -3.62, p = .003^{**}$; *H1e*). However, the number of justified diagnoses in Friction was only marginally higher than in baseline in our test ($t(15) = -1.93, p = .073$). We believe that this might be a result of the relative ease of extracting diagnoses from feedback regardless of tool support for novices, as compared to the more challenging task of developing actionable strategies. Lastly, we analyzed participants' self-reported satisfaction levels with their crafted

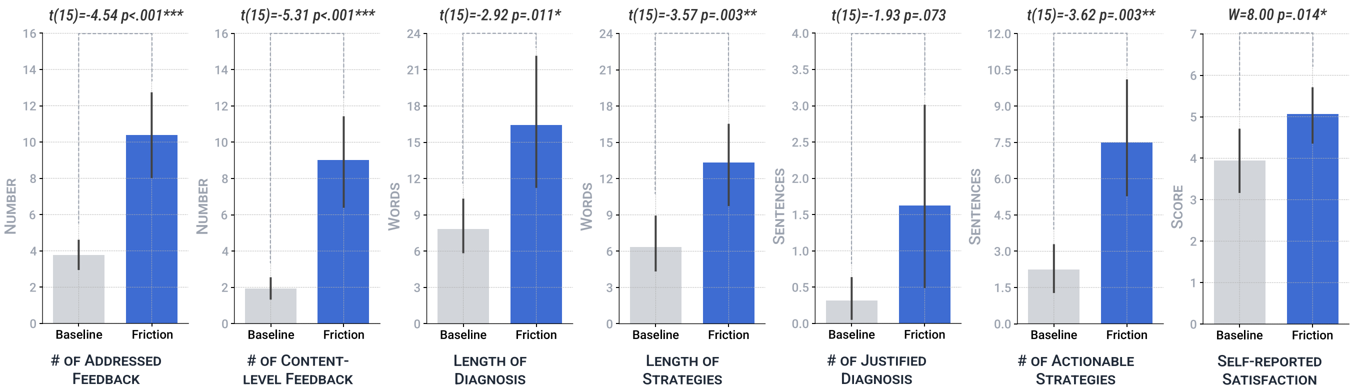

Figure 6: Bar plots illustrating the statistical metrics of participant performance of reflective planning in two conditions, where the t-values from the Student's paired t-test, W-values from the Wilcoxon signed-rank paired test, and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported. Error bars represent 95% confidence intervals (CIs). A table of these values is also provided in the Appendix C.4.1.

reflection plans. The results from a Wilcoxon signed-rank test indicated that participants were significantly more satisfied with the plans they made using Friction ($W = 8.00, p = .014^{*}$; *H1f*).

Based on our observations of the analysis results, we believe that Friction helped participants produce longer, more actionable, and more satisfactory reflection plans.

Qualitative Findings. During reflective planning, Friction aided participants in drilling down to the specifics of each troubled sentence. AI-generated hints played a crucial role at this stage, supporting participants to diagnose specific issues within a sentence and plan revision actions from obscure, vague feedback. For example, P11 mentioned that the AI-generated diagnosis hints "*provided an opportunity to re-examine the writing problem*" and helped them to "*identify precise reasoning issues that were not immediately clear from the initial feedback.*" Importantly, the AI strategy hints further "helped to elaborate on the identified problems" (P01) because they provided "*diverse*" and "*comprehensive*" (P01, P08, and P12) ways of solving a writing issue and "*considered the context*" (P03).

However, in the baseline, when creating revision plans most participants often copied key words or phrases directly from the feedback and pasted them into the text box. This behavior indicated a relatively superficial level of reflection. P11 even noted, "*(the behavior) was due to it (the baseline) workflow requiring me to write a reflective diagnosis and strategies; if it didn't require it, I would skip it and directly ask ChatGPT to help me revise.*"

5.2 Iterative Revision Performance

Participants now move on from planning to conducting their revisions. To evaluate participants' performance in this stage of the process, we first examined the number of revised sentences and the number of iterations per revised sentence. We initially predicted that users of Friction would revise more sentences (*H2a*). However, as shown in Figure 7, we could not detect an observable difference between Friction and baseline in improving the quantity of revised sentences ($t(15) = 0.51, p = .615$). We further calculated the word changes between each pair of original and final revised sentences and also found no observable difference ($t(15) = 0.51, p = .615$; *H2b*). While this breaks with our initial expectations, we noted that a well-performing revision tool could also provide benefits in the form of increased quantity of iterations. While participants spent similar time on revision in both conditions (Table 2), they iterated 0.5 more times per sentence when using Friction than using the baseline ($t(15) = -3.48, p = .003^{**}$; *H2c*).

Our prior results indicate that quantity and speed of revision did not differ, but Friction helped users to iterate might have led to differences in the quality of their changes. To measure this, we calculated and compared the average scores of the revisions rated by two experts in two conditions (outlined in §4.2.4). Experts rated revisions made with Friction as significantly higher compared to the baseline ($t(15) = -2.13, p = .038^{*}$; *H2d*). Moreover, the results from a Wilcoxon signed-rank test showed that participants were significantly more satisfied with the revisions made with Friction ($W = 10.50, p = .046^{*}$; *H2e*), as shown in Figure 7.

Based on our observations of the analysis results, we believe that the increased iterations that Friction facilitated allowed the participants to develop higher quality and more satisfied results.

Qualitative Findings. The AI evaluation played a crucial role in motivating participants to make iterative revisions. Participants like P11 expressed a sense of achievement with positive results, stating, "*I felt a sense of achievement every time I got a green (positive) report. I wanted to turn every sentence into green.*" Curiosity about the explanation led participants like P04 and P16 to try out different versions of revision to probe the evaluation reports. However, their decision to either adopt or reject suggestions depended on how well the report aligned with their expectations. Some sought affirmation from the evaluation, as P13 noted that they "*gained confidence on reflective plans when their revision was evaluated to be better.*" This confirmed belief encouraged bolder refinement. In addition, the evaluation consistently fostered awareness of progress, encouraging participants to "*rethink*" their changes and make informed decisions on iterations. P10, for example, described how the evaluation "*(tells) me whether I'm making progress, so I can*

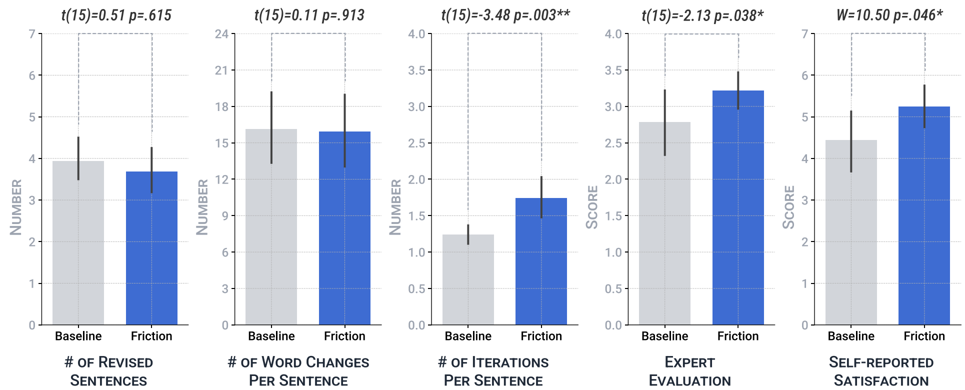

Figure 7: Bar plots illustrating the statistical metrics of participant performance of iterative revision under two conditions, where the t-values from the Student's paired t-test, W-values from the Wilcoxon signed-rank paired test, and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported. Error bars represent 95% confidence intervals (CIs). A table of these values is also provided in the Appendix C.4.1.

decide what I should do next." Initially, they tackled an evidence-based problem with a general statement like "*Research shows...,*" but when the report indicated that the response was still too broad, they replaced it with more specific research examples. This revision led to an improvement, demonstrating how the evaluation guided their decision-making process and iterative refinements.

In the baseline condition, the most common pattern among participants was to input both the feedback snippets and the original sentence, then ask ChatGPT to "*revise the sentence based on the feedback.*" Some participants also provided the entire essay as context. Their iteration process involved first asking ChatGPT to regenerate text, then fine-tuning specific words within the generation.

5.3 Engagement Patterns

Engagement in reflective activities contributes to a learner's artifact improvement and skill development [49, 86, 113]. As shown in Table 1, participants scored Friction significantly higher than the baseline on all four dimensions of UES (FA: $W = 3.00, p = .003^{**}$; PU: $W = 19.00, p = .021^{*}$; AE: $W = 4.50, p = .002^{**}$; RW: $W = 6.00, p = .001^{***}$). For example, they reported being significantly more absorbed with Friction and found it to be more rewarding.

We then took a closer view of the engagement patterns of participants. We examined the timestamps when participants entered each stage of the baseline and Friction, calculating the time spent at each stage. As shown in Table 2, participants using Friction spent a statistically significant 14.58% less time in feedback navigation ($t(15) = 3.95, p = .001^{**}$), while spending a statistically significant 12.35% more time in reflective planning ($t(15) = -2.64, p = .019^{*}$), during a 20-minute task session. A detailed illustration of each participant's time allocation can be checked in Appendix C.4.2. Based on the results, we believe that Friction eased the initial feedback processing, thus freed up more time for participants to reflect.

Qualitative Findings. The significantly reduced time spent on feedback navigation initially demonstrated that the scale of feedback units did not overwhelm users with the help of Friction. Qualitative insights further highlight the role of feedback heatmap in this process. Participants found the feedback heatmap particularly helpful in forming an initial impression of their writing issues, identifying more problematic sections, and noticing the most prominent categories. For instance, it helped P05 "*prioritize revising more troubled sentences in body paragraphs*" and assisted P15 in noticing evidence-related writing problems that they usually overlooked. This approach allowed participants to focus on feedback units of greatest interest, by locating sentences and categories efficiently, mitigating feelings of overload by the entire collection of feedback.

In our quantitative results, we also found that users of Friction spent more time on reflective planning. This is likely because Friction fosters process-oriented learning by requiring active engagement with feedback, unlike the passive acceptance of automated revisions typical of other AI tools. For instance, P03 and P15 noted that Friction encourages "*thinking,*" a critical step often bypassed when AI offers direct corrections. As P15 put it, "*One advantage of Friction over ChatGPT is that it allows me to stop and think about the feedback and its revisions. Otherwise, I directly use the output from ChatGPT as long as it's better than the original sentence.*" We saw other evidence of critical thinking in response to different scaffold modes. P04 shared "*Friction provided hints instead of direct answers, allowing me to learn from examples and practice on my own. However, ChatGPT gave direct revisions. I couldn't learn as effectively merely from the results.*"

5.4 AI Hint Usage Patterns

We examined the AI hint usage behaviors of participants. For the 84 diagnosis-strategy pairs created by participants with Friction, we categorized their creation into three categories: whether the participants did not check AI hints at all, got inspiration from AI hints, or directly adopted AI hints. Although there is no observable difference between participants' usage of AI hints for diagnosis versus strategies, participants directly adopted marginally more hints from AI when devising strategies (Diagnosis: $M = 2.38, SD = 2.85$; Strategies: $M = 2.94, SD = 3.26$). This suggested reliance might be attributed to the more difficult nature of developing strategies compared

Table 1: Survey results of perceived experience on engagement, creativity support, overall usefulness, and task workload under two conditions, where the Wilcoxon signed-rank paired t-test W-values and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

Standardized Scales		Baseline		FRICTION		Statistics	
		M	*SD*	*M*	*SD*	*W*	*p*
User Engagement Scale	Focused Attention	3.25	1.27	4.27	1.03	3.00	.003**
	Perceived Usability	2.54	1.46	3.50	1.68	19.00	.021*
	Aesthetic Appeal	3.19	1.21	5.04	1.23	4.50	.002**
	Reward	3.35	1.51	5.25	1.20	6.00	.001***
Creativity Support Index	Enjoyment	2.66	1.27	4.47	1.58	2.50	.002**
	Exploration	3.47	1.34	5.19	1.11	8.00	.002**
	Expressiveness	2.91	1.23	4.22	1.29	4.00	.006**
	Immersion	2.78	1.05	3.31	1.34	25.50	.094
	Results Worth Effort	3.38	1.20	4.69	1.28	6.50	.004**
	Collaboration	2.47	1.35	4.97	1.41	8.00	.002**
Perceived Overall Usefulness	—	3.00	1.32	4.94	1.65	1.50	.003**
NASA Task Load Index	Mental	5.00	1.71	5.25	1.07	27.50	.642
	Physical	4.31	1.99	3.50	1.79	42.00	.151
	Temporal	4.19	1.42	4.38	1.46	34.50	.751
	Effort	4.44	1.63	4.38	1.20	48.00	.887
	Performance	3.88	1.50	4.81	1.05	17.00	.089
	Frustration	4.25	1.84	3.88	1.75	43.50	.353

Table 2: The statistical metrics of participants' time allocation under two conditions, where the Student's paired t-test t-values and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

Time Allocation	Baseline		FRICTION		Statistics	
	M	*SD*	*M*	*SD*	*t(15)*	*p*
% of Time Spent on Feedback Navigation	38.83	11.42	24.26	14.23	3.95	.001***
% of Time Spent on Reflective Planning	25.84	9.60	38.19	14.66	−2.64	.019*
% of Time Spent on Iterative Revision	35.33	14.36	37.56	12.09	−0.74	.473

to extracting problems, which is also evidenced in §5.1. The detailed data of participant AI hint usage is shown in Appendix C.4.3.

Qualitative Findings. While both conditions included deep engagement with a generative AI model, we found that most participants reported a stronger sense of ownership in their content in FRICTION compared to baseline. For instance, P09 highlighted that FRICTION respected manual effort, stating, "*the final revision was made by myself... FRICTION helped my thinking, but didn't directly change any of my words.*" However, P10, who heavily relied on ChatGPT and felt minimal ownership, noted "*ChatGPT was taking over everything*." Similarly, P03 described ChatGPT as a "*time saver that quickly completed revision on my behalf*," and P09 noted that "*ChatGPT helped with quick revisions*" that they will directly use.

With FRICTION, Participants' usage patterns of AI hints shaped their perception of ownership. Those like P13, P14, and P16, who initially relied on AI hints but gradually reduced their use, reported the highest levels of ownership. For example, P16, evolving from initial AI reliance to independently creating revision plans, reported a highest score of 7. In contrast, others (P01-03, P05-12) who tended to combine AI hints with their own insights felt a strong, yet slightly lesser, sense of ownership. For example, P12, who scored a 5, appreciated the AI's role in shaping direction without restricting creativity, commenting, "*AI hints are very instructive by only giving me a general direction without giving a specific answer. This leaves room for creativity.*" Conversely, an outlier P04, who replicated nearly all AI hints from FRICTION, experienced a comparatively lower sense of ownership (3).

5.5 Perceived Creativity Support

As displayed in Table 1, participants perceived significantly more support from FRICTION than baseline in terms of enjoyment, exploration, expressiveness, and results worth effort. Notably, we changed the original survey questions about collaboration between users to collaboration between the user and AI. The results showed that FRICTION was perceived as easier to collaborate with AI than the baseline ($W = 8.00, p = .002^{**}$). However, there is no observable difference between the two conditions in terms of immersion ($W = 25.50, p = .094$) and both scores are relatively low. We suppose that this was caused by the mentally demanding nature of reflection & revision tasks for novice writers.

5.6 Perceived Usefulness

As shown in Table 1, participants perceived Friction as significantly more useful than the baseline ($W = 1.50, p = .003^{**}$). The top three most useful features were AI hint ($M = 5.81, SD = 1.72$), feedback heatmap ($M = 5.75, SD = 1.06$), and AI evaluation ($M = 5.31, SD = 1.66$). This result aligns with our expectations, as these features were among the most novel and specifically designed for Friction to address key user needs. We illustrate the usefulness ranking of Friction's features in the Appendix C.4.4.

5.7 Perceived Task Workload

We conducted a Wilcoxon signed-rank test to compare participants' perceived workload in two conditions. As shown in Table 1, there is no observable difference between Friction and baseline in all dimensions of NASA Task Load Index. This result suggests that the increased complexity of Friction did not overwhelm users.

5.8 Additional Participant Observations

Participants appreciated the sentence-based mapping between feedback and essays: it preserved context (P15), provided a manageable granularity for novices (P16), and enhanced reflection on specific details (P13). However, some participants (e.g., P06, P15) highlighted a trade-off: certain feedback necessitates revisions at the paragraph, essay, or even topic level. While this larger range of revision can be achieved by revising multiple sentences, a more flexible revision span may achieve a more engaging authoring experience. P02 suggested adding a slider to control the granularity of revision spans, ranging from words and phrases to sentences and paragraphs.

In addition, some participants (e.g., P04, P16) observed that Friction presented a steeper learning curve than the baseline tool, as it required them to learn and adapt to new features such as understanding the helpfulness metrics. The need to engage deeply in reflection also contributed to an increased mental load (e.g., P04, P11), as participants had to "*push themselves to carefully interpret feedback, read suggestions, and make decisions*" (P10). P14 suggested that automatically condensing feedback content and making suggestions more concise could help reduce this load. Participants also reported usability challenges, including issues with the legacy effect when dragging feedback cards, the smoothness of scrolling, and the time spent waiting for AI responses.

Lastly, participants provided valuable recommendations for expanding the system's capabilities. P13 recommended a sorting function to rank feedback cards by the chosen helpfulness metric. P16 proposed adding a feature to automatically cluster feedback units based on common issues. Several participants (e.g., P07, P09, P11, P13) believed Friction could also be useful in other writing scenarios, such as narratives, cover letters, and statements of purpose. P06 suggested the system has potential for training young scholars in academic writing, but P11 cautioned against using it for revising research papers, as "*it needs academic knowledge, statistic data, and innovative ideas*," but AI may not be fully reliable.

6 Discussion

In this paper, we propose and evaluate Friction, a novel AI-infused tool that strategically scaffolds novice writers in structured, in-depth reflection on feedback for writing revision. Based on our findings, we suggest several design implications for future human-AI interaction and creativity support tools.

6.1 Balancing the Trade-Off Between Efficiency and Reflection

The HCI community is currently debating the rapid transformation of the creative industry driven by LLMs [27]. LLMs enable the generation of various forms of text at unprecedented speeds [110], promising to reshape the creative landscape [5]. For instance, writers of all skill levels can now use tools like ChatGPT to produce large volumes of text at their fingertips. HCI researchers have also developed revision tools that generate multiple alternatives in seconds [11, 32, 60–62, 87, 95], enabling less confident writers to easily replace their original work. However, creativity is a nuanced process that often requires sustained focus, time for reflection, and the development of domain-specific skills through long-term dedication [28]—particularly for novices. This creates an urgent need to critically assess the temporal dynamics of LLM systems and explore ways to align their rapid output with the slower, deliberate nature of human creativity to ensure users can develop essential skills through interaction.

Our work leverages AI to provide guidance when participants reflect on their writing issues and potential solutions during the planning phase, aligning with the principles of "cognitive apprenticeship" [19, 20]. In cognitive apprenticeship, a more experienced individual deliberately makes tacit thinking processes visible to the apprentice, enabling the learner to observe, enact, and practice the processes with guidance from the more experienced mentor. Compared to most current AI writing tools—which directly generate revised text for users, bypassing the thinking process—our apprenticeship-based interaction emphasizes making both the thinking process (§3.1) and the AI's step-by-step guidance visible to learners. This approach lowers the barriers to engaging in reflective activities, increases the time participants spend on reflective planning (§5.3), and likely contributed to the observed improvements in participants' performance of planning (§5.1) revision (§5.2).

While we cannot directly measure learning in a single usability study, the observed improvements in behaviors provide preliminary evidence of potential long-term learning gains. According to the theory of deliberate practice [4, 35, 71], sustained focus on challenging tasks, coupled with feedback, leads to the gradual acquisition of expertise. Reflective behavior, emphasized in our work, is a cornerstone of deliberate practice, as it allows individuals to internalize skills through focused effort and self-assessment. Guided reflection through Friction allowed users to move beyond surface-level changes, fostering deeper engagement with the content and helping them understand the nuances of writing issues. However, a trade-off emerges: when AI support focuses primarily on fostering reflection rather than directly assisting users in content generation, efficiency may remain constant or even decline slightly, as evidenced by the number of revised sentences and word changes (§5.2). This work calls for a broader conversation in human-AI interaction on the balance between productivity and reflection, and efficiency and deliberation.

The philosophy of Friction draws from design traditions that push back against the pervasive focus on efficiency above all other metrics in interactive systems—such as slow technology [50, 51], reflective design [102], and design frictions [24, 34]. Cognitive psychology provides further insight into this balance. Kahneman's work [69] describes two modes of thinking: System 1 and System 2. System 1 operates quickly and automatically, guiding routine, mindless behaviors—such as the feedback processing phase. In contrast, System 2 is slower, more deliberate, activated during mindful, reflective tasks like the planning stage in our work. In our work, FRICTION augmented sensemaking during the initial feedback processing, where efficiency is crucial (System 1), while deliberately slowing down the action-planning phase, which requires deeper reflection (System 2). We argue that combining these two modes of thinking could and should be combined in design to optimize the use of LLMs in creative processes. Future endeavors could explore how to seamlessly integrate both systems in LLM-enhanced tools in broader creative domains (e.g., visual design).

6.2 Leveraging Visualization as a Proxy for Complex AI Outputs

LLM outputs are purely textual and often unstructured, which can be challenging to interpret at scale. In our work, textual feedback comments are organized within an interactive heatmap, spread across different sections of the essay and various types of writing problems. This visualization proved effective in guiding user attention toward problematic content and salient issues (§5.1): Participants used it to prioritize areas needing revision and to systematically address a specific type of issues in different sections. This aligns with prior research [56–58], which showed that visualizations are powerful interfaces for representing complex AI reasoning and uncovering hidden patterns in human-AI co-writing.

As LLMs become increasingly integrated into creative and intellectual tasks, the need for intuitive and scalable methods of interaction is all the more pressing [46]. In our work, when the input (i.e., feedback) was large in volume, the output of LLMs (i.e., predictions of problematic sentences) expanded proportionally, further complicating user sensemaking. In other words, when assistance was most needed, it became paradoxically harder and harder to interpret models' assistance. The interactive heatmap emerges as a suitable proxy to represent LLM output and effectively bridges the gap between feedback and writing. By highlighting areas of concern, users were guided toward a more systematic workflow that balances addressing both local and global issues during revision. This work extends prior research, demonstrating the critical role of visualizing AI outputs in directing human attention and augmenting user sensemaking.

However, while visualizing LLM output may help to reduce complexity and draw user attention, reducing complexity by its nature is lossy. Some amount of useful context may be abridged by the visualization, diminishing potential benefits. For users of FRICTION, we aimed to strike the right balance between providing summaries in visual form and access to raw text when desired, but there is no canonical strategy for doing so. It remains to be seen where and when this approach is most advantageous, and whether it scales smoothly with different levels of task and feedback complexity. In addition, future studies could also investigate how different types of visualizations and varying levels of interactivity in these visualizations would affect human sensemaking of LLM outputs.

6.3 Preserving Ownership While Being Scaffolded by AIs

Our study provides initial evidence that offering participants hints to scaffold their revision efforts, rather than replacing their work, can enhance their sense of ownership (§5.4). While participants generally reported a strong sense of ownership, this was influenced by how they used AI-generated hints. Specifically, a greater reliance on AI hints was linked to a reduced feeling of ownership. This aligns with prior research showing that increased AI involvement in text generation—from none to sentence-level to paragraph-level contributions—diminishes individuals' sense of ownership over the text [30]. In addition, Zhou and Sterman [134] suggested that imperfect, intermediate AI-generated text, which leaves room for writers to modify, may foster greater ownership compared to fluent AI continuations. A similar dynamic may be at play when an AI offers hints instead of full text generation.

Prior work showed that timing of feedback delivery can impact recipients' feedback-seeking behaviors and creative outcomes [33]. This is similar to timing of hint provision. For instance, giving students more agency in the planning process—by providing AI-generated hints after they have written an initial draft themselves—could act as a "second opinion" and potentially bolster their sense of ownership [103]. In addition, Aleven et al. [3] found that adaptive fading of hints led to greater robust learning than fixed-fading and no fading conditions. We can imagine that if an AI was able to monitor user performance in reflection and dynamically adjust the granularity and frequency of their hints, both learning effectiveness and sense of ownership might be further enhanced. On the other hand, it may not be the presence of hints which improves ownership, but rather agency in terms of activating and using the system.

In light of these initial findings, we encourage future researchers to explore several key areas: comparing creators' sense of ownership when AI provides scaffolding in the form of hints versus full text generation, examining the relationship between user engagement with AI hints and their perception of ownership, and investigating the impact of hint timing and fading on creators' sense of ownership. We believe that these future explorations would offer valuable insights for designing appropriate AI scaffolds that preserve ownership while providing desired creativity support.

7 Limitations

There are a few limitations to consider with regard to our prototype system, study methodology, and analysis. First, since we chose argumentative writing as an example domain to help design the system, the prompting method for feedback categorization is designed specifically for this genre. Creative genres may differ substantially in the issue topics addressed by critics. For example, instead of claims, reasoning and evidences, feedback for fictions may be related to character, settings, and story lines. However, we believe that the design of workflows, visualizations, and AI scaffolds is applicable across domains.

Second, our participants provided valuable suggestions for improving and expanding the system. We are working on implementing these suggestions to make FRICTION more usable. Some suggestions also open avenues for future research. For instance, we followed previous work to adopt a sentence-based revision model given its advantages of representing the natural boundary of text and proved effectiveness in supporting revisions [1, 60, 111, 130]. However, sentence-based revision may not apply to all scenarios. This raises a broader question about the most appropriate level of granularity for feedback. While FRICTION provided an initial exploration in this area, future study is warranted to understand how adjusting the scope of revisions shapes the process.

Third, with regard to our methodology, we followed previous work [122] to provide participants with existing essays and feedback in the tasks. They might react differently if they were revising their own work based on the feedback collected by themselves. For example, a user might reflect more on writing habits when reviewing salient issues from the heatmap. Moreover, we collected feedback through crowdsourcing. Previous research has shown that crowd workers could potentially use LLMs to complete tasks [109]. We adopted several methods from prior work [108] to prevent LLM use, as detailed in Appendix B.1. However, we acknowledge that the potential for LLM use by crowd workers cannot be entirely eliminated. Additionally, while we adopted computational linguistic methods from prior work to quantitatively assess reflection plans, we recognize that reflection is inherently nuanced and complex, which may not be fully captured by these computational techniques. Future research could benefit from a more detailed, qualitative content analysis to better understand the differences in the nature of plans developed with and without FRICTION.

Lastly, the size of the user study ($N = 16$) is also a limitation. A more extensive study in the future, involving a larger and more diverse participant pool, would provide a stronger foundation for evaluating the system's usability and generalizability across different user groups, including more experienced writers and professional writing contexts. For example, we can analyze how different demographic factors (e.g., gender, age, writing proficiency, AI tools used, frequency, and purposes) may influence the quantitative results tested in the study. Moreover, our study's duration was based on prior research and preliminary investigations, revealing initial evidence of participants' more engaged behaviors and the positive immediate outcomes from it. To measure learning effect and writing skill development, we plan to conduct a long-term, longitudinal study.

8 Conclusion

In conclusion, our work with FRICTION highlights the potential of generative AI in transforming the feedback-driven revision process for novice writers by striking a critical balance between efficiency and reflection. By strategically breaking down and organizing feedback, FRICTION encourages deeper engagement with content-related issues, leading to more thoughtful and effective revisions. The integration of adaptive hints and feedback heatmap further aids in navigating complex feedback, enhancing sensemaking, and promoting iterative reflection. As AI continues to evolve, its role in creativity should not merely be to simplify tasks but to improve learning experience, enabling users to develop greater mastery and creative autonomy. Our findings open new avenues for the design of human-AI systems that promote deliberate practice and reflection, offering a promising direction for future research in creativity support and beyond.

Acknowledgments

We sincerely thank our participants for sharing their thoughts and feedback on our system, as well as the reviewers for their valuable insights and suggestions.

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. https://doi.org/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. https://doi.org/10.18653/v1/W18-0528

[3] Vincent Aleven, Ido Roll, Bruce M. McLAREN, and Kenneth R. Koedinger. 2010. Automated, Unobtrusive, Action-by-Action Assessment of Self-Regulation During Learning With an Intelligent Tutoring System. *Educational Psychologist* 45, 4 (Oct. 2010), 224–233. https://doi.org/10.1080/00461520.2010.517740

[4] Joanna Allan. 1996. Learning Outcomes in Higher Education. *Studies in Higher Education* 21, 1 (Jan. 1996), 93–108. https://doi.org/10.1080/03075079612331381487

[5] Nantheera Anantrasirichai and David Bull. 2022. Artificial Intelligence in the Creative Industries: A Review. *Artif Intell Rev* 55, 1 (Jan. 2022), 589–656. https://doi.org/10.1007/s10462-021-10039-7

[6] 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. https://doi.org/10.3200/TCHS.81.4.149-154

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

[8] Bernadette Berardi-Coletta, Linda S. Buyer, Roger L. Dominowski, and Elizabeth R. Rellinger. 1995. Metacognition and Problem Solving: A Process-Oriented Approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21, 1 (1995), 205–223. https://doi.org/10.1037/0278-7393.21.1.205

[9] Elizabeth L. Bjork and Robert A. Bjork. 2011. Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning. *Psychology and the real world: Essays illustrating fundamental contributions to society* 2, 59-68 (2011), 56–64.

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

[11] 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. https://doi.org/10.1145/3411764.3445372

[12] David Stephen Calonne. 2006. Creative Writers and Revision. *Revision: History, theory, and practice* 1 (2006), 142–176.

[13] 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. https://doi.org/10.1145/3415232

[14] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.* 21, 4 (Aug. 2014), 1–25. https://doi.org/10.1145/2617588

[15] Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting Self-Explanations Improves Understanding. *Cognitive science* 18, 3 (1994), 439–477.

[16] 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]

[17] Heeryung Choi, Jelena Jovanovic, Oleksandra Poquet, Christopher Brooks, Srećko Joksimović, and Joseph Jay Williams. 2023. The Benefit of Reflection Prompts for Encouraging Learning with Hints in an Online Programming Course. *The Internet and Higher Education* 58 (June 2023), 100903. https://doi.org/10.1016/j.iheduc.2023.100903

[18] 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. https://doi.org/10.1145/3491102.3501819

[19] Allan Collins, John Seely Brown, and Ann Holum. 1991. Cognitive Apprenticeship: Making Thinking Visible. *American educator* 15, 3 (1991), 6–11.

[20] Allan Collins and Manu Kapur. 2006. *Cognitive Apprenticeship*. Vol. 291. na, na.

[21] Amy Cook, Steven Dow, and Jessica Hammer. 2020. Designing Interactive Scaffolds to Encourage Reflection on Peer Feedback. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1143–1153. https://doi.org/10.1145/3357236.3395480

[22] 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. https://doi.org/10.1145/3290605.3300368

[23] National Research Council. 2000. *How People Learn: Brain, Mind, Experience, and School* (expanded edition ed.). National Academies Press, Washington, D.C. https://doi.org/10.17226/9853

[24] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[25] 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. https://doi.org/10.1145/3571817

[26] Patrick A. Crain and Brian P. Bailey. 2017. Share Once or Share Often? Exploring How Designers Approach Iteration in a Large Online Community. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. Association for Computing Machinery, New York, NY, USA, 80–92. https://doi.org/10.1145/3059454.3059476

[27] 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. https://doi.org/10.1145/3643834.3660688

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

[29] E.A. Davis. 2003. Prompting Middle School Science Students for Productive Reflection: Generic and Directed Prompts. *Journal of the Learning Sciences* 12, 1 (2003), 91–142. https://doi.org/10.1207/S15327809JLS1201_4

[30] 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]

[31] Jim Dillard and Cindy Harmon-Jones. 2002. A Cognitive Dissonance Theory Perspective on Persuasion. In *The Persuasion Handbook: Developments in Theory and Practice*. SAGE Publications, Inc., Washington DC, 99.

[32] 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. https://doi.org/10.18653/v1/2022.acl-long.250 arXiv:2203.03802 [cs]

[33] Jane L. E, Yu-Chun Grace Yen, Isabelle Yan Pan, Grace Lin, Mingyi Li, Hyoungwook Jin, Mengyi Chen, Haijun Xia, and Steven P. Dow. 2024. When to Give Feedback: Exploring Tradeoffs in the Timing of Design Feedback. In *Creativity and Cognition*. ACM, Chicago IL USA, 292–310. https://doi.org/10.1145/3635636.3656183

[34] Jonathan Ericson. 2023. Reimagining the Role of Friction in Experience Design. *J. User Exper.* 17, 4 (June 2023), 131–139.

[35] K. Anders Ericsson. 2006. The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, New York, NY, US, 683–703. https://doi.org/10.1017/CBO9780511816796.038

[36] 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. https://doi.org/10.1145/3301275.3302265

[37] Ralph P. Ferretti, William E. Lewis, and Scott Andrews-Weckerly. 2009. Do Goals Affect the Structure of Students' Argumentative Writing Strategies? *Journal of Educational Psychology* 101, 3 (2009), 577–589. https://doi.org/10.1037/a0014702

[38] Ralph P. Ferretti, Charles A. MacArthur, and Nancy S. Dowdy. 2000. The Effects of an Elaborated Goal on the Persuasive Writing of Students with Learning Disabilities and Their Normally Achieving Peers. *Journal of Educational Psychology* 92, 4 (2000), 694–702. https://doi.org/10.1037/0022-0663.92.4.694

[39] Jill Fitzgerald and Lynda R. Markham. 1987. Teaching Children About Revision in Writing. *Cognition and Instruction* 4, 1 (March 1987), 3–24. https://doi.org/10.1207/s1532690xci0401_1

[40] Matthew Fledderjohann. 2019. *When Writers Encounter Dissonance*. Ph.D. Dissertation. The University of Wisconsin - Madison, United States – Wisconsin.

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

[42] 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. https://doi.org/10.2307/357381 jstor:357381

[43] 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. https://doi.org/10.1145/3134680

[44] 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]

[45] 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. https://doi.org/10.1145/3532106.3533533

[46] 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. https://doi.org/10.1145/3613904.3642139

[47] Steve Graham, Karen R. Harris, Mary Adkins, and April Camping. 2021. Do Content Revising Goals Change the Revising Behavior and Story Writing of Fourth Grade Students At-Risk for Writing Difficulties? *Read Writ* 34, 7 (Sept. 2021), 1915–1941. https://doi.org/10.1007/s11145-021-10142-9

[48] 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. https://doi.org/10.1145/2757226.2757249

[49] 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. https://doi.org/10.1145/3544548.3581054

[50] Lars Hallnäs. 2015. On the Philosophy of Slow Technology - Acta Universitatis Sapientiae. https://acta.sapientia.ro/en/series/social-analysis/publications-acta-social/social-analysis-contents-of-volume-5-no-1-2015/on-the-philosophy-of-slow-technology.

[51] Lars Hallnäs and Johan Redström. 2001. Slow Technology – Designing for Reflection. *Personal Ub Comp* 5, 3 (Aug. 2001), 201–212. https://doi.org/10.1007/PL00000019

[52] Karen Handley, Margaret Price, and Jill Millar. 2011. Beyond 'Doing Time': Investigating the Concept of Student Engagement with Feedback. *Oxford Review of Education* 37, 4 (Aug. 2011), 543–560. https://doi.org/10.1080/03054985.2011.604951

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

[54] 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. https://doi.org/10.1016/S0166-4115(08)62386-9

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

[56] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1260–1276. https://doi.org/10.1145/3532106.3533526

[57] Md Naimul Hoque, Bhavya Ghai, Kari Kraus, and Niklas Elmqvist. 2023. Portrayal: Leveraging NLP and Visualization for Analyzing Fictional Characters. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 74–94. https://doi.org/10.1145/3563657.3596000

[58] 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. https://doi.org/10.1145/3613904.3641895

- [59] 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. https://doi.org/10.1145/3272973.3274069
- [60] Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. 2020. Langsmith: An Interactive Academic Text Revision System. arXiv:2010.04332 [cs]
- [61] 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]
- [62] 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. https://doi.org/10.1145/3586183.3606810
- [63] Maria Jackson and Leah Marks. 2016. Improving the Effectiveness of Feedback by Use of Assessed Reflections and Withholding of Grades. *Assessment & Evaluation in Higher Education* 41, 4 (May 2016), 532–547. https://doi.org/10.1080/02602938.2015.1030588
- [64] 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. https://doi.org/10.1145/3544548.3581196
- [65] Thorben Jansen, Jennifer Meyer, Johanna Fleckenstein, Andrea Horbach, Stefan Keller, and Jens Möller. 2024. Individualizing Goal-Setting Interventions Using Automated Writing Evaluation to Support Secondary School Students' Text Revisions. *Learning and Instruction* 89 (Feb. 2024), 101847. https://doi.org/10.1016/j.learninstruc.2023.101847
- [66] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (March 2023), 248:1–248:38. https://doi.org/10.1145/3571730
- [67] Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the Human Revision Process in Scientific Writing. arXiv:2210.15067 [cs]
- [68] Anders Jonsson. 2013. Facilitating Productive Use of Feedback in Higher Education. *Active Learning in Higher Education* 14, 1 (March 2013), 63–76. https://doi.org/10.1177/1469787412467125
- [69] Daniel Kahneman. 2013. *Thinking, Fast and Slow* (first edition ed.). Farrar, Straus and Giroux, New York.
- [70] 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. https://doi.org/10.1145/3173574.3174180
- [71] RONALD T. KELLOGG and ALISON P. WHITEFORD. 2009. Training Advanced Writing Skills: The Case for Deliberate Practice. *Educational Psychologist* 44, 4 (Oct. 2009), 250–266. https://doi.org/10.1080/00461520903213600
- [72] 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. https://doi.org/10.1145/2998181.2998196
- [73] 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. https://doi.org/10.1145/3613904.3642529 arXiv:2403.00439 [cs]
- [74] Stephen King. 2000. *On Writing: A Memoir of the Craft.* Simon and Schuster, New York, NY, USA.
- [75] 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. https://doi.org/10.1145/3025453.3025883
- [76] 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. https://doi.org/10.1080/14703297.2019.1687005
- [77] 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. https://doi.org/10.1111/bjet.13054
- [78] 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. https://doi.org/10.1145/3613904.3642697 arXiv:2403.14117 [cs]
- [79] 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. https://doi.org/10.1145/3491102.3502030
- [80] 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. https://doi.org/10.18653/v1/2022.in2writing-1.9
- [81] Äli Leijen, Kai Valtna, Djuddah A.J. Leijen, and Margus Pedaste. 2012. How to Determine the Quality of Students' Reflections? *Studies in Higher Education* 37, 2 (March 2012), 203–217. https://doi.org/10.1080/03075079.2010.504814
- [82] Teresa Limpo, Rui A. Alves, and Raquel Fidalgo. 2014. Children's High-level Writing Skills: Development of Planning and Revising and Their Contribution to Writing Quality. *Brit J of Edu Psychol* 84, 2 (June 2014), 177–193. https://doi.org/10.1111/bjep.12020
- [83] Xiaodong Lin and James D. Lehman. 1999. Supporting Learning of Variable Control in a Computer-Based Biology Environment: Effects of Prompting College Students to Reflect on Their Own Thinking. *Journal of Research in Science Teaching* 36, 7 (1999), 837–858. https://doi.org/10.1002/(SICI)1098-2736(199909)36:7<837::AID-TEA6>3.0.CO;2-U
- [84] 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. https://doi.org/10.1145/2675133.2675283
- [85] Charles A. MacArthur. 2016. Instruction in Evaluation and Revision. In *Handbook of Writing Research, 2nd Ed.* The Guilford Press, New York, NY, US, 272–287.
- [86] Jennifer Marlow and Laura Dabbish. 2014. From Rookie to All-Star: Professional Development in a Graphic Design Social Networking Site. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 922–933. https://doi.org/10.1145/2531602.2531651
- [87] 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]
- [88] 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. https://doi.org/10.1145/3544548.3581023
- [89] 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.
- [90] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* 112 (April 2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004
- [91] Diane Pecher and Rolf A. Zwaan. 2005. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking.* Cambridge University Press, Cambridge, England.
- [92] Adam M. Persky and Jennifer D. Robinson. 2017. Moving from Novice to Expertise and Its Implications for Instruction. *Am J Pharm Educ* 81, 9 (Nov. 2017), 6065. https://doi.org/10.5688/ajpe6065
- [93] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. National Institute of Standards and Technology, McLean, VA, USA, 2–4.
- [94] Margaret Price, Karen Handley, Jill Millar, and Berry O'Donovan. 2010. Feedback : All That Effort, but What Is the Effect? *Assessment & Evaluation in Higher Education* 35, 3 (May 2010), 277–289. https://doi.org/10.1080/02602930903541007
- [95] 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. https://doi.org/10.48550/arXiv.2310.00117 arXiv:2310.00117 [cs]
- [96] Thomas P. Ryan and J. P. Morgan. 2007. Modern Experimental Design. *Journal of Statistical Theory and Practice* 1, 3-4 (Dec. 2007), 501–506. https://doi.org/10.

1080/15598608.2007.10411855
[97] Joan M. Sargeant, Karen V. Mann, Cees P. van der Vleuten, and Job F. Metsemakers. 2009. Reflection: A Link between Receiving and Using Assessment Feedback. *Adv in Health Sci Educ* 14, 3 (Aug. 2009), 399–410. https://doi.org/10.1007/s10459-008-9124-4
[98] 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. https://doi.org/10.1145/3450741.3465253
[99] Donald A. Schon. 2010. Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions. *Australian Journal of Adult Learning* 50, 2 (2010), 448–451.
[100] Donald A. Schön. 2017. *The Reflective Practitioner: How Professionals Think in Action*. Routledge, London. https://doi.org/10.4324/9781315237473
[101] Raymond Scupin. 1997. The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization* 56, 2 (1997), 233–237. jstor:44126786
[102] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. 2005. Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility (CC '05)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569
[103] Anjali Singh, Christopher Brooks, Xu Wang, Warren Li, Juho Kim, and Deepti Wilson. 2024. Bridging Learnersourcing and AI: Exploring the Dynamics of Student-AI Collaborative Feedback Generation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. Association for Computing Machinery, New York, NY, USA, 742–748. https://doi.org/10.1145/3636555.3636853
[104] 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.
[105] Sarah Sterman, Evey Huang, Vivian Liu, and Eric Paulos. 2020. Interacting with Literary Style through Computational Tools. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376730
[106] Chad C. Tossell, Nathan L. Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J. De Visser. 2024. Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. *IEEE Transactions on Learning Technologies* 17 (2024), 1069–1081. https://doi.org/10.1109/TLT.2024.3355015
[107] 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. https://doi.org/10.1145/3170427.3188474
[108] 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. https://doi.org/10.48550/arXiv.2310.15683 arXiv:2310.15683
[109] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. https://doi.org/10.48550/arXiv.2306.07899 arXiv:2306.07899
[110] 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. https://doi.org/10.1002/jocb.597
[111] 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. https://doi.org/10.1145/3313831.3376732
[112] Bruce Weigl. 1976. Revision as a Creative Process. *English Journal* 65, 6 (Sept. 1976), 67–68. https://doi.org/10.58680/ej197614763
[113] 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. https://doi.org/10.1080/00461520.2016.1207538
[114] 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. https://doi.org/10.1080/03075079.2015.1130032
[115] Carol Withey. 2013. Feedback Engagement: Forcing Feed-Forward amongst Law Students. *The Law Teacher* 47, 3 (Dec. 2013), 319–344. https://doi.org/10.1080/03069400.2013.851336
[116] 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. https://doi.org/10.1145/3491102.3517582
[117] 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. https://doi.org/10.1145/3555210
[118] 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. https://doi.org/10.1145/2531602.2531604
[119] Xiaotong (Tone) Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L. E, and Steven P. Dow. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 907–921. https://doi.org/10.1145/3640543.3645196
[120] 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. https://doi.org/10.1145/2901790.2901820
[121] 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. https://doi.org/10.1145/3059454.3059468
[122] 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. https://doi.org/10.1145/3313831.3376380
[123] 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. https://doi.org/10.1145/3490099.3511105
[124] 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. https://doi.org/10.1145/2818048.2819953
[125] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3613904.3642647
[126] Chao Zhang, Cheng Yao, Jianhui Liu, Zili Zhou, Weilin Zhang, Lijuan Liu, Fangtian Ying, Yijun Zhao, and Guanyun Wang. 2021. StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–6.
[127] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501914
[128] Chao Zhang, Zili Zhou, Yajing Hu, Lanjing Liu, Jiayi Wu, Yaping Shao, Jianhui Liu, Lingyan Zhang, Lijuan Liu, Hangyue Chen, Fangtian Ying, and Cheng Yao. 2023. Observe It, Draw It: Scaffolding Children's Observations of Plant Biodiversity with an Interactive Drawing Tool. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference (IDC '23)*. Association for Computing Machinery, New York, NY, USA, 253–266. https://doi.org/10.1145/3585088.3589380
[129] 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. https://doi.org/10.18653/v1/P17-1144
[130] 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. https://doi.org/10.18653/v1/N16-3008
[131] Fan Zhang and Diane Litman. 2015. Annotation and Classification of Argumentative Writing Revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Joel Tetreault, Jill Burstein, and Claudia Leacock (Eds.). Association for Computational Linguistics, Denver, Colorado, 133–143. https://doi.org/10.3115/v1/W15-0616
[132] 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. https://doi.org/10.1145/3586183.3606800
[133] David Zhou and Sarah Sterman. 2023. Creative Struggle: Arguing for the Value of Difficulty in Supporting Ownership and Self-Expression in Creative Writing. In *The Second Workshop on Intelligent and Interactive Writing Assistants (In2Writing)*. In2Writing, New York, NY, USA, 1–3.
[134] David Zhou and Sarah Sterman. 2024. Ai.Llude: Investigating Rewriting AI-Generated Text to Support Creative Expression. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*. Association for Computing Machinery, New York, NY, USA, 241–254. https://doi.org/10.1145/3635636.3656187
[135] Wei Zhu. 2001. Performing Argumentative Writing in English: Difficulties, Processes, and Strategies. *TESL Canada Journal* 19, 1 (2001), 34–50.

A GPT-4 Prompts Used in FRICTION

Here, we outline the techniques we employed to guide the GPT-4 model within the context of FRICTION. We followed the work of Wu et al. [116] to break down complex tasks into a set of LLM primitive operations and borrowed their proposed default prompting and data structures to design instructions. The temperature of the model is set to 0.7 for creative operations like Strategy Generation and 0.0 for factual or deterministic tasks like Feedback Classification.

A.1 Feedback Segmentation

System Prompt:
Split the feedback paragraph into an ordered list of different writing problems. Group adjacent sentences that target the same problem or subject. Keep the original text. No rephrasing. Correct mis-spelled words in the feedback. Ignore sentences that are not stating problems. Return 0 if there is no problem provided.

....

Examples:
Input:
feedback: Egad! First of all, the rubric related to grammar is so hideously abused here it is hard to get over. There, I have said that, and I will not linger on it, but really? The structure and organization is not bad. The student stated the issue, let us know they disagreed with the premises, broke their argument down into useful subsets, and addressed what they thought the issue was and why they believed it was possible to overcome each issue. They also provided a conclusion. This is the only plus, but it should have some weight, since so many cannot muster it. The issues are poorly described, and the reason behind the issue misunderstood. There are far more salient points to make on each of the sub-topics, and only a little research would have helped to see them. The suggestions for how to overcome the perceived issues were laissez-faire, generic and not proven or supported by any discussion of merit. As for fluidity, I suppose despite being largely oblique and unsupported, the essay enjoys a reasonable flow, and one poor point is followed smoothly to another poor point. The language precision is dreadful. I would love to edit this with this student and push them to work a little harder and research much more, since the bones are there. All word processors contain very easy to use spelling and grammar functions, and the student would be wise to spend the several minutes required to take advantage of them.
Output:
1. First of all, the rubric related to grammar is so hideously abused here it is hard to get over.
2. The issues are poorly described, and the reason behind the issue misunderstood. There are far more salient points to make on each of the sub-topics, and only a little research would have helped to see them.
3. The suggestions for how to overcome the perceived issues were laissez-faire, generic and not proven or supported by any discussion of merit.
4. As for fluidity, I suppose despite being largely oblique and unsupported, the essay enjoys a reasonable flow, and one poor point is followed smoothly to another poor point.
5. The language precision is dreadful.
6. All word processors contain very easy to use spelling and grammar functions, and the student would be wise to spend the several minutes required to take advantage of them.
Input:
feedback: The essay contains grammatical errors. The sentence "some students utilize their cell phones as an effective cheating tools..." should not have the word "an." The sentence "the mobile phone has become a dispensable multifunctional tools" should have singular "tool" instead of "tools." The writer gives specific examples to support their argument, such as when they explained how mobile phones can cause social issues. However, some of their points are vague and seem rushed. For example, they claim that the social problems and privacy concerns of mobile phone users can be solved, but they don't give any ideas on how to solve them. The writer doesn't cite any sources to support this claim: "It has been proved that overusing of the electronic devices including mobile phones could lead to higher possibility of suffering hearing loss and even cancers, although the further investigation are needed."
Output:
1. The essay contains grammatical errors. The sentence "some students utilize their cell phones as an effective cheating tools..." should not have the word "an."
2. The sentence "the mobile phone has become a dispensable multifunctional tools" should have singular "tool" instead of "tools."
3. However, some of their points are vague and seem rushed. For example, they claim that the social problems and privacy concerns of mobile phone users can be solved, but they don't give any ideas on how to solve them.
4. The writer doesn't cite any sources to support this claim: "It has been proved that overusing of the electronic devices including mobile phones could lead to higher possibility of suffering hearing loss and even cancers, although the further investigation are needed."

A.2 Feedback Classification

The prompt that we used to classify the purposes of feedback consists of two steps. We first instruct the model to categorize the given feedback into "Surface" or "Content," and then tasked it with a detailed taxonomy.

System Prompt:
Step-1: Categorize the given feedback into "Surface" or "Content" purposes. Surface feedback focuses on critiquing the language, organization, and technical aspects of the writing, including word usage, grammar, fluency, and sentence structure. Content feedback critiques the substance of the essay, including the claims, reasoning, evidence, and counterarguments presented. Respond with the most relevant category.

Step-2 if surface: Categorize the given feedback into "Organization", "Word-usage" or "Orthography" purposes. Organization feedback helps the author get a better flow of the paper. Word-usage feedback critiques words or phrases for better representation of ideas. Orthography feedback focuses on spelling or grammar errors, misusage of punctuation, or adherence to the organizational conventions of academic writing. Respond with the most relevant category.

Step-2 if content: Categorize the given feedback into "Claim", "Reasoning", "Evidence", "Rebuttal", or "Others" purposes. Claim feedback critiques the position or claim being argued for. Reasoning feedback focuses on the principle or reasoning of the claim. Evidence feedback critiques facts, theorems, or citations for supporting claims/ideas. Rebuttal feedback focuses on the development of content that rebuts current claims/ideas. Others critique content that does not directly support or rebut claims/ideas. Respond with the most relevant category.

....

Examples:
Input:
feedback: I also don't understand why the author did not clearly state a main idea to start out the essay, such as I am here today to convince you of why cooperation is more important to a child's development as opposed to competing rather than to wait until the end.
Output:
Step-1: Content
Step-2: Claim
Input:
feedback: However, I would also argue that this argumentative essay lacks quantitative data to support these claims. The author should be able to identify specific economic data to be able to prove that the impact of international tourism grows the local

economy.
Output:
Step-1: Content
Step-2: Evidence
Input:
feedback: I just do not feel that the writer talked about how cruel it is to animals to be tested on and how it makes them feel. Many people love animals so I can see how they would not like what the writer wrote.
Output:
Step-1: Content
Step-2: Reasoning
Input:
feedback: The essay does not appear to provide any counter examples or arguments, and opposing positions are only briefly acknowledged in the first paragraph without any expansion on the arguments.
Output:
Step-1: Content
Step-2: Rebuttal
Input:
feedback: The conclusion felt brief and did not tie up many loose ends. It felt like it was the briefest possible summary.
Output:
Step-1: Content
Step-2: Others
Input:
feedback: Also, some of the work choices were awkward, such as 'borne in mind' and 'serious problems springing from modern technology.' Again, this is a simple fix for the author.
Output:
Step-1: Surface
Step-2: Word-usage
Input:
feedback: The claims are easy to understand but I did not like the structure of the body paragraphs and felt it was not clear enough.
Output:
Step-1: Surface
Step-2: Organization
Input:
feedback: The sentence "the mobile phone has become a dispensable multifunctional tools" should have singular "tool" instead of "tools."
Output:
Step-1: Surface
Step-2: Orthography

A.3 Problematic Sentence Detection and Diagnosis

We direct the model to adopt the role of an "*English writing teacher*" to set the tone. The original full essay will be injected as context information. The output is required to be structured in the JSON format for information extraction.

System Prompt:
You are an English writing teacher. Your job is to detect a list of sentences from the essay to address the given feedback. If the feedback suggests revising certain sentences, return the sentence index and give a specific and justifiable rationale to diagnose why they need to be revised rather than tell how to revise them. If the feedback suggests adding new content, identify the certain sentence index after which the new content should be inserted and explain why it should be positioned there.
Only respond in JSON structure like this: ["id": <index start from 1>, "text": "<original text>", "rationale": "<rationale>"]
Essay: injecting the full essay content
....

Examples:
Input:
feedback: However, there is no clear evidence cited in his essay response, only vague generalities and "more studies are needed" responses.
Output:
["id": 10, "text": "It has been proved that overusing of the electronic devices including mobile phones could lead to higher possibility of suffering hearing loss and even cancers, although the further investigation are needed.", "rationale": "This sentence makes a claim about the health effects of overusing electronic devices, but does not provide any specific evidence or sources to support this claim. It also ends with a vague statement about needing further investigation.",
"id": 13, "text": "This possibility exists all the time. But a more efficient and reliable system can also be invented to guarantee the private information of users.", "rationale": "This sentence makes a claim about the possibility of conversations being intercepted and the potential for a more reliable system, but does not provide any specific evidence or sources to support these claims."]
Input:
feedback: In terms of language usage, the essay contains several grammatical errors and awkward phrasings. For example, the phrase "Nowadays, the popularity of mobile phones has brought about a lot of convenience but at the meanwhile a variety of problems as well" could be revised for clarity and conciseness.
Output:
["id": 1, "text": "Nowadays, the popularity of mobile phones has brought about a lot of convenience but at the meanwhile a variety of problems as well", "rationale": "The phrase 'at the meanwhile' is grammatically incorrect and awkward. The correct phrase should be 'in the meantime' or 'meanwhile.' Additionally, the sentence is quite long and could be made more concise.",
"id": 5, "text": "Mobile phones are not out of expectation.", "rationale": "The phrase 'not out of expectation' is awkward and unclear. It's not immediately clear what the writer means by this.",
"id": 7, "text": "What is worse, some students utilize their cell phones as an effective cheating tools in examinations in which they can send and receive answers by texting each other.", "rationale": "The phrase 'as an effective cheating tools' is grammatically incorrect. The singular form 'tool' should be used instead of the plural 'tools."',
"id": 8, "text": "Nevertheless, it is not impossible to solve these problems, as long as some regulations and rules can be effectively implemented.", "rationale": "The phrase 'it is not impossible to solve these problems' is a double negative and could be rephrased for clarity.",
"id": 9, "text": "It is also worth mentioning that some harmful effects on our health are lethal.", "rationale": "The phrase 'some harmful effects on our health are lethal' is awkward and could be rephrased for clarity.",
"id": 14, "text": "Moreover, the mobile phone has become a dispensable multifunctional tools in people's daily life.", "rationale": "The phrase 'a dispensable multifunctional tools' is grammatically incorrect. The singular form 'tool' should be used instead of the plural 'tools."']

A.4 Revision Strategy Generation

System Prompt:
You are an English writing teacher. Your job is to offer three one-sentence, diverse, specific, and actionable strategies for revising the given sentence in the essay below, based on the provided feedback. Do not include any revised sentences in your response.
Only respond in JSON structure like this: ["title": <three word strategy name>, "strategy": "<strategy content>"]
Essay: injecting the full essay content
....

Examples:
Input:
feedback: Using a quote that, I also had to infer was a quote because it was stated like a quote but wasn't quoted nor cited is not much support for convincing one of their side of an argument.
sentence: It is always said that competition can effectively promote the development of economy.
Output:
[“title”: “Cite Source Clearly”, “strategy”: “Explicitly mention the source of the quote about competition promoting economic development to add credibility to the argument.”,
“title”: “Avoid Unattributed Quotes”, “strategy”: “Rephrase the sentence about competition promoting economic development without using an unattributed quote to strengthen the argument.”,
“title”: “Provide Supporting Evidence”, “strategy”: “Include factual data or examples to support the claim that competition is effective in promoting economic development.”]
Input:
feedback: I also do not think the conclusion does a good job of going through the main points of why a cooperative attitude is superior than a competitive one. There are no new insights in the conclusion.
sentence: Consequently, no matter from the view of individual development or the relationship between competition and cooperation we can receive the same conclusion that a more cooperative attitudes towards life is more profitable in one's success.
Output:
[“title”: “Summarize Main Points”, “strategy”: “Summarize the key points discussed in the essay regarding the benefits of cooperation over competition to strengthen the conclusion.”,
“title”: “Include New Insights”, “strategy”: “Add fresh insights or perspectives in the conclusion that were not explicitly mentioned in the body of the essay to make the conclusion more impactful.”,
“title”: “Provide Actionable Steps”, “strategy”: “Offer specific actions or recommendations based on the advantages of a cooperative attitude to make the conclusion more practical and engaging.”]

A.5 Plan Implementation Analysis

The prompt that we used to provide analysis on the plan implementation of each iteration is depend on the results of our fine-tuned GPT-3 model's results about revision improvements (seen in §3.4).

System Prompt:
If not better: You are an English writing teacher. Your job is to analyze how the student failed in using the following strategies in revision and therefore, makes no improvement to the original sentence. Keep your analysis within 30 words. Start with “The revision ...”
Strategies: injecting the action plans crafted by the user
If better: You are an English writing teacher. Your job is to analyze how the student succeeded in using the following strategies in revision and therefore, makes improvements to the original sentence. Keep your analysis within 30 words. Start with “The revision ...”
Strategies: injecting the action plans crafted by the user
....
Examples:
Input:
original sentence: In addition, hi-tech devices also have a very positive impact on many aspects of our lives.
revised sentence: Also, high-tech devices are really good for healthcare, like because of a 30% increase in recovery, like with better diagnostic things and monitoring patients more efficiently.
Output:
The revision lacks a clear presentation of statistics, using vague language (”really good” and “like”) instead of precise terms. The phrase “30% increase in recovery” lacks context, and “better diagnostic things” is informal.
Input:
original sentence: Personally, I partly disagree with this point of view, and the reasons are given below.
revised sentence: I disagree with this view that we should refrain from technology advancement, due to the vast opportunities it enables despite potential dangers.
Output:
The revised sentence provides a clear statement of disagreement, specifies the topic, and hints at the reasons for disagreement making it more informative and engaging. It also avoids unnecessary phrases like “personally” and “the reasons are given below”, making it more concise and direct.

B Technical Evaluation

B.1 Process of Crafting Feedback Dataset

We first collected a corpus of feedback following human annotation to establish a ground truth. To create this corpus, we sampled six argumentative essays from a widely used dataset [104]. This dataset consists of 90 argumentative essays collected from an online community where English as a Second Language (ESL) learners share essays they write to practice for standardized English tests like IELTS. We chose this dataset because the authors of the papers match the target users of our system, i.e., novices. The sampled essays covered different topics, such as education, technology, economy, and policy. For each essay, we recruited 10 crowd workers to provide feedback. We also provided a rubric to guide the feedback generation process, adopted from Harrell's work [53]. This approach aligns with previous research indicating that crowdsourced feedback, especially when guided by a rubric to enhance its quality, is comparable in terms of quality, scope, and length to that generated by social media, online communities, and even experts [120, 121, 124]. Each feedback task paid $4.

Previous research has shown that crowd workers could potentially use LLMs to complete tasks [109]. To prevent LLM use, we followed the work of Veselovsky et al. [108] by explicitly instructing crowd workers not to use LLMs and converting the original rubric into an image. However, we acknowledge that even with both methods, we cannot entirely eliminate the potential for LLM use by crowd workers. Given that the quality of collected feedback is rated as relatively high by trained expert evaluators in Section 4.2.2 and considering the previous usage of AI in providing feedback on writing content [7], we believe that occasional instances of LLM usage in the feedback data would not significantly affect data quality.

One research assistant divided the 60 pieces of feedback into 223 unique feedback items, each targeting one issue in the associated essay. We sampled 100 items as our final corpus for subsequent annotation and evaluation.

B.2 Additional Comments on Performance of LLM Pipelines in Detection, Diagnosis, and Strategies

B.2.1 Comments on LLM Performance in Detection. Evaluators commented that our LLM method effectively scanned the entire essay for issues, not only catching problematic sentences in

Table 3: Demographic information of participants and their usage of AI writing tools in daily writing practice. *Writing proficiency. **English proficiency (both on Likert scales from 1/low to 7/high). ***Usage purposes of AI writing tools include: 1—Grammar and spelling assistance, 2—Writing prompt or idea suggestions, 3—Refining writing styles or tones, 4—Polishing content or language, 5—Plagiarism detection, 6—Summarizing text, 7—Reflecting on writing, 8—Understanding feedback.

ID	Gender	Age	Writing*	English**	AI Tool Types	AI Usage Freq.	AI Usage Purpose***
P01	Female	23	4	4	ChatGPT	Weekly	1, 2, 3, 4, 6
P02	Female	22	4	4	ChatGPT, Grammarly	Daily	1, 3, 4
P03	Female	24	2	4	ChatGPT, Grammarly, Notion AI	Daily	1, 2, 3, 7
P04	Male	26	2	3	ChatGPT, Grammarly, Notion AI	Weekly	1, 3, 4
P05	Female	22	4	6	ChatGPT	Weekly	3, 4
P06	Female	25	3	3	ChatGPT, Grammarly	Weekly	3, 4, 6
P07	Female	24	4	5	ChatGPT, Gemini	Weekly	2, 4
P08	Female	35	4	5	ChatGPT, Grammarly, Deeplearn	Daily	1, 2, 3, 4
P09	Female	23	3	5	ChatGPT, Grammarly	Weekly	1, 2, 3, 6
P10	Male	22	4	6	ChatGPT	Daily	1, 2, 3, 4, 6, 9
P11	Female	24	3	5	ChatGPT	Weekly	1, 2, 4, 6
P12	Male	—	4	4	ChatGPT	Weekly	2, 3
P13	Male	28	3	5	ChatGPT, Grammarly	Weekly	1, 3, 4
P14	Male	24	3	7	ChatGPT, Bing Chat, Gemini	Monthly	7
P15	Male	23	4	5	ChatGPT, DeepL	Daily	1, 3, 4
P16	Male	23	3	3	ChatGPT, Grammarly, Notion AI	Weekly	2, 3, 4, 6

noticeable areas like the beginning paragraph but also thoroughly reviewing the body paragraphs. It was particularly adept at identifying multiple problematic sentences simultaneously, maintaining accuracy without the risk of fatigue. However, the LLM sometimes failed to identify all problematic sentences when the feedback was overly specific. For instance, when feedback highlighted specific sentences that needed revision, the LLM sometimes neglected to scan for additional sentences in other areas of the essay.

B.2.2 Comments on LLM Performance in Diagnosis. Evaluator feedback indicated that LLM-generated diagnoses were generally objective, straightforward, and concise. They focused on logical explanations without directly proposing specific solutions, aligning with expectations for the diagnosis phase. These diagnoses avoided emotional or ambiguous expressions, instead favoring declarative statements that are easily understood by novices. For instance, LLM-generated diagnoses clearly point out issues within the context of both feedback and writing: "*It is unclear what strategies or methods can be used to encourage scientists to pay more attention to developing new technologies.*"

While these statements were clear, the evaluators suggested that the diagnosis would have been more thought-provoking if it had taken a questioning tone. For example, "*How do you address counterarguments against international tourism?*," could inspire deeper reflection compared to declarative statements like "*this sentence does not discuss any counterarguments.*" Additionally, there were edge cases where the LLM misinterpreted the author's intent, likely due to context ambiguity or conflicting counterarguments.

B.2.3 Comments on LLM Performance in Strategies. In the strategizing phase, evaluator believed that our method excelled in generating creative and diverse strategies, offering distinct perspectives on problem-solving. For example, when optimizing a quote, the LLM suggested not only to "*include a source to clarify its origin*" but also to "*paraphrase the statement instead of presenting it as a quote.*" Our LLM method was particularly good at using complete and detailed examples when describing strategies, especially for evidence or reasoning issues. For instance, when given feedback that "*the writer does not go into specifics about how tourism helps locals with jobs*," rather than suggesting a vague "*be more specific,*" the LLM provided actionable strategies: "*elaborate on the specific benefits of these tourism-related jobs, such as higher wages, better working conditions, or job security compared to other local jobs.*"

Similar to diagnosis, the strategies generated from our method were sensitive to the context of feedback, sentences, and instructions. However, the LLM sometimes failed to incorporate distant context, leading to strategies that were either too broad or redundant. For instance, when the LLM overlooked the overarching theme or existing examples, their suggestions often lacked relevance or added unnecessary repetition.

C User Evaluation

C.1 Participant Information

The demographic information of participants is shown in Table 3.

C.2 Interview Protocol

These are the questions used in the semi-structured interview after the two revision sessions.

(1) What was your overall experience with each tool? Was there anything that excited or frustrated you?
(2) Comparing the baseline and Friction, what were the main differences you noticed in the feedback-driven revision process?

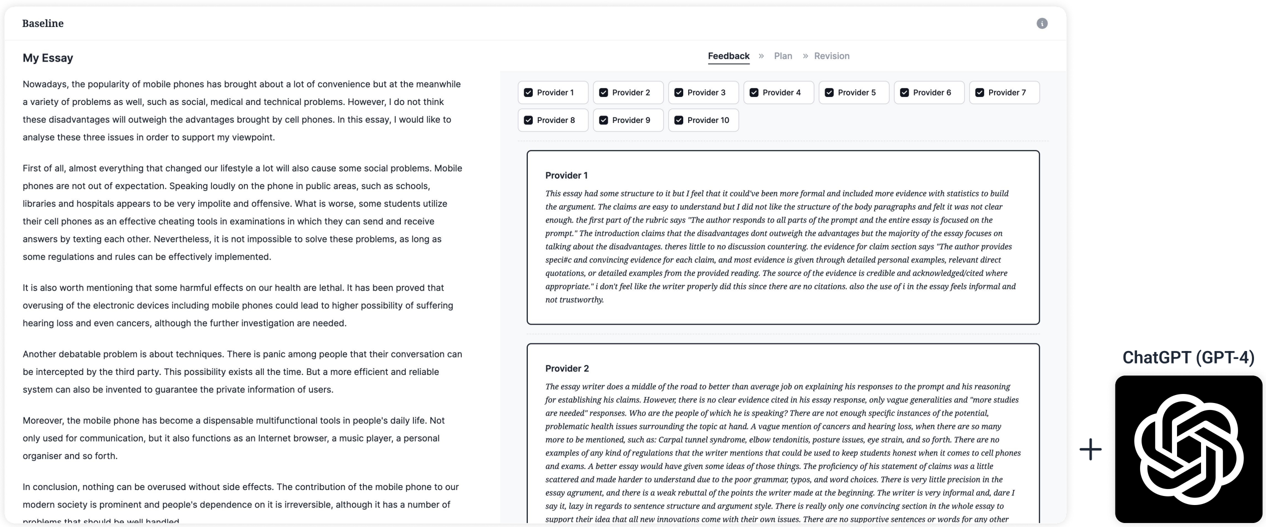

Figure 8: Screenshot of the baseline system.

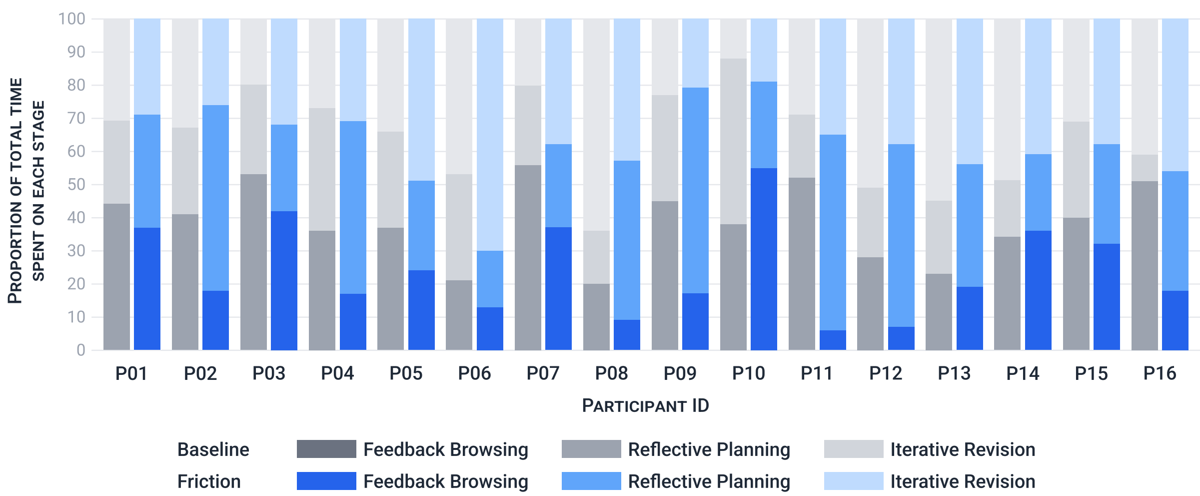

Figure 9: Participants' time distribution across three stages in a 20-minute task session: *Feedback Browsing*, *Reflective Planning*, and *Iterative Revision*. The time spent on reflective planning is marked with dashed borders.

(3) In the three main stages of feedback-driven revision—feedback browsing, reflective planning, and iterative revision—did you find one tool to be more helpful than the other, and why?
(4) Were there any differences in your typical approach to feedback-driven revision when using these tools? If so, how was it different from your usual work flow?
(5) Which features did you find most beneficial in both tools, and in what scenarios were they particularly useful?
(6) Do you think the two tools will contribute to your long-term writing skill development? If so, how?
(7) How much ownership do you feel for what you created in the revision task? Rate from 1 to 7 for both tools and explain why.
(8) What role do you think the two tools play in your feedback-driven revision?
(9) Do you have any suggestions or ideas to improve the tools?

C.3 Baseline System Interface

The baseline system shared a similar interface with Friction but without its major advancements. In baseline, users could manually select feedback pieces, reflect on the selected feedback, and make necessary revisions on certain sentences. To ensure improvements were due to Friction's unique features and not general AI capabilities, participants in baseline also had access to ChatGPT[11], a widely-used generative AI tool for writing assistance [106].

[11] https://chat.openai.com/

Table 4: The statistical metrics of participant performance of reflective planning, where the t-values from the Student's paired t-test, W-values from the Wilcoxon signed-rank paired test, and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

Reflective Planning Performance		Baseline		Friction		Statistics			Hypotheses
		M	*SD*	*M*	*SD*	$t(15)$	*W*	*p*	
Addressed Feedback	# of Addressed Feedback	3.75	1.65	10.38	5.23	−4.54	—	.001***	*H1a* supported
	# of Content-level Feedback	1.94	1.18	9.00	5.16	−5.31	—	.001***	*H1b* supported
Reflective Plan	Length of Diagnosis	7.82	4.88	16.44	11.24	−2.92	—	.011*	*H1c* supported
	Length of Strategies	6.33	4.81	13.33	6.73	−3.57	—	.003**	
	# of Justified Diagnosis	0.31	0.60	1.63	2.68	−1.93	—	.073	*H1d* marginal
	# of Actionable Strategies	2.25	2.15	7.50	5.19	−3.62	—	.003**	*H1e* supported
	Self-reported Satisfaction	3.94	1.61	5.06	1.39	—	8.00	.014*	*H1f* supported

Table 5: The statistical metrics of participant performance of iterative revision, where the t-values from the Student's paired t-test, W-values from the Wilcoxon signed-rank paired test, and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

Iterative Revision Performance	Baseline		Friction		Statistics			Hypotheses
	M	*SD*	*M*	*SD*	$t(15)$	*W*	*p*	
# of Revised Sentences	3.93	1.06	3.69	1.14	0.51	—	.615	*H2a* not supported
# of Word Changes Per Sentence	16.14	6.02	15.93	6.69	0.11	—	.913	*H2b* not supported
# of Iterations Per Sentence	1.24	0.27	1.74	0.58	−3.48	—	.003**	*H2c* supported
Expert Evaluation	2.78	1.19	3.22	0.88	−2.13	—	.038*	*H2d* supported
Self-reported Satisfaction	4.44	1.50	5.25	1.13	—	10.50	.046*	*H2e* supported

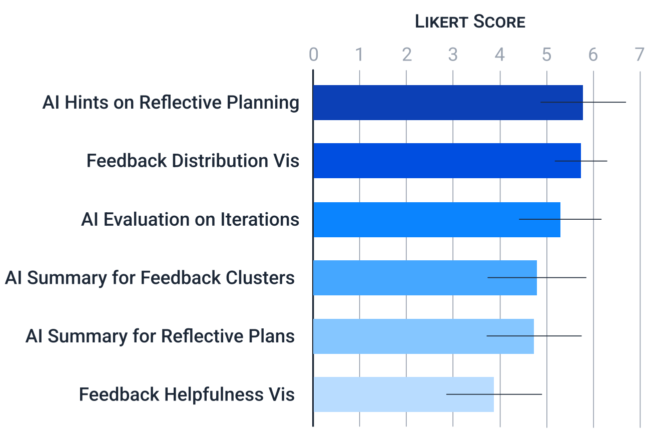

Figure 10: Bar plot of perceived usefulness. Error bars represent 95% confidence intervals (CIs) of mean.

C.4 Additional Results

C.4.1 Task Performance. Table 4 and 5 show detailed statistical results of participant performance in reflective planning and iterative revision separately, including the results from Student's paired t-tests and Wilcoxon signed-rank paired tests.

C.4.2 Time Distribution. We investigated each participant's time allocation in detail during the feedback-driven revision tasks in both conditions. As shown in Figure 9, Friction increased the time spent in reflective planning for 12 participants (except P03, P05, P06, and P10), while helping 14 participants (except P10 and P14) spend less time in feedback browsing.

Table 6: The statistical metrics of participants' average AI hint usage, where the Student's paired samples t-test t-values and p-values (*: p<.05, **: p<.01, ***: p<.001) are reported.

	Diagnosis		Strategies		Statistics	
	M	*SD*	*M*	*SD*	*t(15)*	*p*
No AI Hint Usage	1.18	1.68	1.31	1.70	−0.40	.697
AI Hint Inspiration	1.69	2.72	1.00	2.00	1.23	.239
AI Hint Adoption	2.38	2.85	2.94	3.26	−1.19	.254

C.4.3 AI Hint Usage. For the 84 diagnosis-strategy pairs created by participants with Friction, we categorized their creation into three categories: whether the participants did not check AI hints at all, got inspiration from AI hints, or directly adopted AI hints. As shown in Table 6, although there is no observable difference between participants' usage of AI hints for diagnosis versus strategies, they directly adopted marginally more hints from AI when devising strategies (Diagnosis: $M = 2.38, SD = 2.85$; Strategies: $M = 2.94, SD = 3.26$).

C.4.4 Perceived Usefulness. We took a detailed look into the usefulness of the individual features of Friction. As shown in Figure 10, The top three most useful features of Friction were AI Hint ($M = 5.81, SD = 1.72$), feedback heatmap ($M = 5.75, SD = 1.06$), and AI Evaluation on Iterations ($M = 5.31, SD = 1.66$). This result aligns with expectations, as these features were among the most novel and specifically designed to address key user needs.