Experiment-Inspired Design: Enhancing Digital Learning Through Parallel Exploration of Multiple Design Alternatives

by

Mohi Reza

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Computer Science
University of Toronto

Experiment-Inspired Design: Enhancing Digital Learning Through
Parallel Exploration of Multiple Design Alternatives

Mohi Reza
Doctor of Philosophy

Department of Computer Science
University of Toronto
2025

Abstract

Despite the flexibility of digital learning—where content updates can be deployed instantly—many digital learning resources such as explanations, motivational messages, emails, and hints are rarely tested against a broad range of alternatives to determine what works best. Current authoring and evaluation workflows emphasize incremental improvements through **linear** updates, offering limited support for systematically comparing design alternatives in **parallel**. Even when educational experiments are conducted, they are often one-off studies focused on proving an idea's efficacy, rather than enabling **perpetual** experimentation, where ideas are continually tested and adapted. As a result, content development tends to prioritize a supposed 'best' version based on limited evidence, rather than embracing a systematic cycle of exploration and refinement.

Two interlinked challenges continue to limit progress toward continuous, experiment-driven improvement: (1) the lack of *authoring workflows* that support parallel creation and management of multiple alternative designs without overwhelming users, and (2) the lack of *deployment mechanisms* for testing those variations in authentic learning environments, automating processes where possible while ensuring human oversight.

To address these challenges, this dissertation proposes **experiment-inspired design**—an approach to improving digital learning resources by creating, testing, and refining multiple content alternatives in parallel. Rather than relying on linear updates or isolated experiments, experiment-inspired design supports rapid cycles of content exploration and refinement, enabling continuous improvement across different kinds of digital learning resources. A key feature of this approach is its emphasis on both quantitative and qualitative assessment of content variations, providing deeper insight into what works, for whom, and why, in real-world learning contexts. It is also pedagogical-framework-agnostic: rather than endorsing a particular instructional theory, it equips educators, instructional designers, and researchers to rapidly generate and test different pedagogical strategies. While experiment-inspired design does not require artificial intelligence, recent advances in generative AI and reinforcement learning make it significantly easier to implement—lowering the barriers to

parallel content authoring and evaluation through automation. This dissertation demonstrates how AI systems can accelerate experiment-inspired design while preserving human agency, supporting human authors in generating, organizing, and evaluating content without ceding control.

To advance experiment-inspired design and directly address the challenges of parallel content authoring and evaluation, this dissertation presents five projects, each contributing new methods and tools for improving digital learning resources.

1. **The Eustress Intervention** (Chapter 2) demonstrates how systematically designing and evaluating content variations can address critical learner challenges and improve learning outcomes in real-world contexts.

2. **The AdapComp Framework** (Chapter 3) formalizes a flexible framework for running continuous experiments in digital learning platforms, unifying experimentation, dynamic improvement, and personalization in digital learning.

3. **The ABScribe Interface** (Chapter 4) introduces a human-AI authoring interface that accelerates the parallel exploration and organization of content variations, streamlining the revision process through parallel editing.

4. **The PromptHive System** (Chapter 5) extends ABScribe's capabilities to a collaborative setting, enabling domain experts to guide AI models in generating and refining content through a human-centered prompt authoring workflow.

5. **Preserving Human Agency** (Chapter 6) presents a systematic review of 109 HCI papers selected from over 1,600 in the human-AI collaborative writing space, showing how generative AI can be integrated into content authoring while maintaining human control and oversight.

Together, these projects introduce a novel approach to educational experimentation, contributing open-source technologies that support continuous improvement and personalization of digital learning. By bridging principles from traditional and adaptive A/B testing, HCI research on parallel prototyping, reinforcement learning, and recent advancements in generative AI, this work empowers educators, instructional designers, and researchers to advance more adaptive, evidence-driven digital learning through perpetual experimentation.

Acknowledgements

This dissertation would not have been possible without the support of my PhD advisor, committee members, mentors, colleagues, friends, and family. I would like to specifically thank:

- My advisor, Joseph Jay Williams, for bringing enthusiasm and energy to my research, and for encouraging me to continually rethink and improve ideas—not only in research but also in life.
- My committee members and external examiners, Tovi Grossman, Olivier St-Cyr, René Kizilcec, Cosmin Munteanu, and Sean Rintel for their insightful feedback, which has greatly sharpened the framing of my research. I could not have asked for a more supportive and helpful committee.
- My mentors and research collaborators, Anastasia Kuzminykh, Zachary A. Pardos, John Stamper, Norman Bier, Juho Kim, Michael Liut, and Andrew Petersen, for taking time from your busy schedules to provide invaluable research and career advice. My graduate experience would not have been as enriching without your guidance.
- Michelle Craig, Jacqueline Smith, Tom Fairgrieve, and Paul Gries, for allowing me to deploy my research in your courses and for encouraging me to improve the quality of my field interventions to ensure a positive impact on students.
- Friends and colleagues at the University of Toronto, Ilya Musabirov, Ananya Bhattacharjee, Harsh Kumar, Tong Li, Angela Zavaleta, Fred Haochen Song, Nathan Laundry, Dana Kulzhabayeva, Karthik Mahadevan, Anna Offenwanger, Rachel Phinnemore, Bryan Wang, Blaine Lewis, Dashiel Carerra, Pan Chen, Bingjian Huang, Fernando Yanez, and others at the DGP lab, for turning what might have been a solitary experience into something far more collaborative and fun.
- Friends and colleagues at CMU LearnLab, Steven James Moore and Marshall An, and at the CAHL Group at UC Berkeley, Ioannis Anastapolous and Shreya Bhandari, for helping me expand the scope and impact of my research across multiple digital learning platforms and institutions.
- My undergraduate advisees, including Peter Dushniku, Yuming Huang, Jessica Lee, Calista Barber, Michael Zhi-Yuan, Jeb Thomas-Mitchell, Zhongyuan Liang, Joe Fang, Sarva Sanjay, and others, for giving me the joy of mentoring.
- Internship mentors at the Human-Machine Interaction Lab at Huawei, Wei Li and Soheil Kianzad, for helping me push the boundaries of my research in an industry setting.
- Dongwook Yoon and Joanna McGrenere, for introducing me to the wonderful field of HCI. Sandra Mathison for cultivating a deep appreciation for qualitative research in me and Ed Kroc for teaching me how to design and analyze quantitative experiments.
- My oldest friends, Aaraf Afzal, Sifana Sohail, Selima Kabir, Sayema Hossain, Aninda Dibya Saha, Syed Tahmid Mahbub, Naveed Hossain, SK Sajidul Kadir, Nishat Shama, Shafaat Mridha, Sakib Matin, Fahimul Huq, and Anika Alam Purba—for being just a phone call away and for their support, laughter, and friendship over the years.
- My brother, whose early guidance sparked my love for Computer Science.
- My parents, for their unwavering love and support.
- Labiba, for being there through thick and thin.

Contents

List of Tables

List of Figures

List of Appendices

Statement of Contributions

This dissertation includes first-authored papers that have been published in, or are currently under review by, peer-reviewed conferences in human-computer interaction. Published papers appeared in the proceedings of the Association for Computing Machinery (ACM), and permission has been granted by the ACM for their inclusion in this thesis. Where applicable, each chapter includes references to the corresponding publications listed below.

Chapter 2: The Eustress Intervention
Mohi Reza, Angela Zavaleta Bernuy, Emmy Liu, Tong Li, Zhongyuan Liang, Calista K Barber, and Joseph Jay Williams. 2023. *Exam Eustress: Designing Brief Online Interventions for Helping Students Identify Positive Aspects of Stress.* In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 439, 1–13. https://doi.org/10.1145/3544548.3581368

Chapter 3: The AdapComp Framework
Mohi Reza, Juho Kim, Ananya Bhattacharjee, Anna N. Rafferty, and Joseph Jay Williams. 2021. *The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses.* In Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21). Association for Computing Machinery, New York, NY, USA, 15–26. https://doi.org/10.1145/3430895.3460128

Chapter 4: The ABScribe Interface
Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. *ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models.* In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1042, 1–18. https://doi.org/10.1145/3613904.3641899

Chapter 5: The PromptHive System
Mohi Reza, Ioannis Anastasopoulos, Shreya Bhandari, and Zachary A. Pardos. 2025. *PromptHive: Bringing Subject Matter Experts Back to the Forefront with Collaborative Prompt Engineering for Educational Content Creation.* In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 148, 1–22. https://doi.org/10.1145/3706598.3714051

Chapter 6: Preserving Human Agency
Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. 2025. *Co-Writing with AI, on Human Terms: Aligning Research with User Demands Across the Writing Process.* Manuscript under review. Preprint available at https://arxiv.org/abs/2504.12488

To Maa, who let me fly. To Labiba, who flew with me.

Chapter 1

Introduction

> Design is about exploring and comparing the relative merits of alternatives.
>
> *Bill Buxton [42]*

What if digital learning felt more like learning from an attuned instructor—one whose explanations, motivational messages, hints, and emails were continually refined to meet learners' ever-evolving needs? Instead of grappling with rigid webpages and files mimicking the limitations of print, delivering the same content to all learners, what if instructors could design resources by testing new ideas as they arise? By systematically comparing new ideas against the relative merits of existing alternatives, they could discover what actually works[1]—and use insights from those comparisons to inspire even better versions. Over the past five years, through conversations with many instructors and from my own experience in teaching and conducting experiments in large flipped classrooms with thousands of students, I have observed first-hand the challenges of trying to replicate the fluid, continuous experimentation that we instinctively do when teaching face-to-face. In live settings, we can adapt explanations on the fly based on students' responses, trying alternative approaches until comprehension *clicks* and their faces light up.

[Motivation (CR3e): Clarified that the fluidity of teaching in-person comes from face-to-face interactivity, which enables instructors to adapt instruction to learner needs. The revised argument reframes adaptability—not authoring effort—as the key challenge, and positions parallel editing as a mechanism to support it in digital contexts.] But in digital learning, those faces remain unseen. The fluidity of in-person instruction arises from face-to-face interactivity, which enables instructors to adapt instruction dynamically in response to learners' needs. Digital content lacks this kind of real-time interactivity, yet it offers greater flexibility in terms of editing and distribution. We can leverage this flexibility by building tools that support analogous adaptability: that is, the ability to easily test, compare, and refine multiple alternative approaches. Doing so requires overcoming two interlinked obstacles: (1) the lack of *authoring workflows* that support the **parallel creation and management of multiple content variations**, and (2) the lack of *deployment mechanisms* to **systematically deliver and evaluate[2] those variations in authentic learning environ-**

[1] About 80 to 90% of education interventions do not work when tested rigorously [298], underscoring the need not just for more experimentation, but for scalable, low-cost workflows that support continuous refinement.

[2] Ideally, evaluations should combine quantitative metrics with systematic *qualitative* assessments that capture

ments.

Combined, these obstacles create a level of complexity that makes continuous experimentation on content variations overly daunting. To foster a culture of continuous experimentation, instructors and other stakeholders in digital learning need interface and infrastructure support that reduces the marginal cost of running an experiment to near zero. Such support could help educators and instructional teams replicate the rapid, data-driven improvement cycles seen in the tech industry, where dedicated experimentation platforms—such as Ax at Meta [16] and ExP at Microsoft [177]—have enabled large-scale, data-driven optimization. By conducting over 20,000 controlled experiments annually [177], major tech companies have discovered simple yet highly effective changes, such as a minor adjustment in a search engine ad display that led to a remarkable 12% increase in revenue [177]. Now, imagine achieving a comparable improvement on learning outcomes through small yet powerful modifications to digital content. A core aim of my doctoral research has been to advance this vision by developing open-source tools that make continuous experimentation on digital learning as fluid as in-person instruction with minimal costs. Reducing the cost of experimentation is not only about financial expense; it is equally about minimizing the cognitive workload for both instructors and students, who already face significant demands on their time and attention. This is why many of the tools introduced in this thesis focus on using AI to reduce task workload without compromising instructional quality.

The benefits of parallel exploration of alternative designs isn't limited to gathering insights from online experiments. In traditional design practice, such as graphic design, industrial design, and architecture, considering multiple alternative designs is the norm and pervades the entire design process [42]. Studies in HCI on parallel prototyping have shown how exploring many alternatives in parallel can yield better design results through increasing the self-efficacy of the designer [79], avoiding fixation on a singular idea [151, 108], reducing chances of eliminating rough but innovative ideas due to premature evaluation [112, 42], and making people less prone to inflated subjective appraisals by giving them an opportunity to critically assess ideas in relation to each other [345, 42]. This implies that considering many alternative designs can be useful in its own right, even when we do not have a large population of users to run quantitative experiments, and that qualitative reflection from co-designers and users could be helping in informing ways to improve the design.

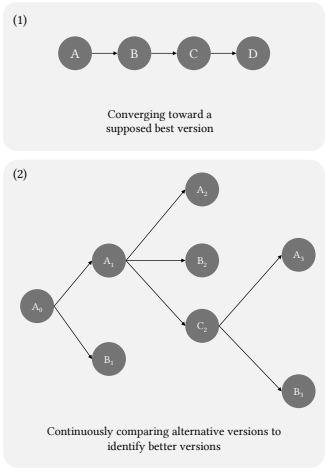


Figure 1.1: Linear progression versus parallel exploration

To explore ways to enable parallel exploration of multiple design alternatives in digital learning, I proposes the concept of *experiment-inspired design*. [Definition (CR4a): Clarified what I mean by "parallel exploration of multiple design alternatives."]By parallel exploration, I am referring to the practice of simultaneously creating and evaluating not just one, but many alternative versions of the learning materials. Each version is a design alternative that encapsulates different ideas from instructors for enhancing the learning materials. There is no "best version" at any given

meaningful differences between design variants while retaining some of the benefits from the structure of controlled experiments [213]. ABScribe and PromptHive, introduced in Chapters 4 and 5 respectively, provide concrete interfaces for integrating this process during content authoring.

time, but rather an ongoing pool of competing ideas that are continually tested and refined through comparisons. This is in contrast to how content is currently managed and improved—typically via linear updates to converge toward a supposed best version without systematically comparing against alternatives that could be better. Figure 1.1 contrasts between the two approaches.

Drawing inspiration from A/B[3] experiments and randomized controlled trials—the gold standard of validating the efficacy of alternatives—as well as HCI theories on parallel prototyping and traditional design, experiment-inspired design serves as useful terminology for encapsulating a broader view of what it means to run an education experiment. Instead of the commonly-held notion that the purpose of an experiment is to either accept or reject a pre-defined hypothesis, it re-positions experimentation as a means to explore the complex design space of creating and improving instructional content.

While these principles guide the design of effective instructional tools, their urgency is underscored by the growing global demand for scalable, high-quality digital learning–an issue highlighted in the United Nations' recent call to action [350]. Today, digital learning plays a critical role in expanding access to education and meeting the urgent needs of learners worldwide, ensuring resilience in the face of disruptions such as the COVID-19 pandemic and future unforeseen challenges. It is no longer an optional component of learning, but essential to our broader goal of providing quality education for all. The UN outlines three key factors for unlocking the transformative potential of digital learning:

1. **Content:** Developing high-quality curriculum-relevant digital educational resources that address learners' needs and interests.

2. **Capacity:** Ensuring instructors, learners, and other stakeholders have the capacity to utilize digital tools for learning through evidence-based approaches.

3. **Connectivity:** Leveraging the internet to connect learners and other stakeholders through the use of digital tools and resources.[4]

1.1 Research Objective: Enhancing Digital Learning

The core objective of this dissertation is to investigate ways to accelerate and improve how we enhance digital learning **content** by developing tools that increase the **capacity** of instructors and stakeholders to explore and evaluate different instructional approaches, while leveraging opportunities arising from increased **connectivity** between digital learning platforms and their users. For example, the shift from offline to online learning opens up the ability to collect data on how learners interact with the materials, and use that to improve and adapt content. Despite these opportunities, digital learning materials are seldom tested against alternatives to determine the most effective approaches and explore better designs. Although freed from the constraints of print, the pace of improvement in digital content remains slow, and changes are often incremental.

The slow pace of innovation isn't due to a lack of ingenuity or effort from instructors, but rather reflective of the limitations of existing content authoring workflows in supporting them with the rapid

[3] Although commonly referred to as A/B testing, the method can involve more than just two conditions.

[4] The UN's focus on this factor emphasizes expanding reliable internet connectivity to a larger population. However, I see this as merely a starting point for what connectivity can enable—such as empowering educators and researchers to discover more effective teaching methods and enhance learning resources through systematic experimentation.

exploration and evaluation of multiple alternative instructional approaches. Every time an instructor has a new idea, it presents an untapped opportunity to leverage digital learning technologies to help them quickly incorporate that idea into content and test it in real-world contexts. However, this is often challenging because instructors are already overburdened with managing the multitude of platforms required to deliver resources to students. To be viable, any tool we design must integrate seamlessly with existing authoring workflows—otherwise, it becomes unrealistic to expect instructors to manage not just one version of content, but many. [Motivation (CR3d): Briefly discussed how "the lack of authoring workflows that support parallel creation and management of multiple content variations without overwhelming users" applies to rich-media course management systems like D2L Brightspace, Moodle and Google Classroom]While rich-media course platforms like Moodle, Google Classroom, and D2L Brightspace do offer flexible content creation tools, they are typically not designed to support structured, *parallel* development and evaluation of multiple content variations. Instead, they subscribe to a linear model where a single "best" version is maintained (as opposed to supporting iterative experimentation on competing alternatives). The tools developed in this dissertation aim to extend and augment these existing workflows by embedding experimentation, comparison, and revision more directly into the authoring process.

Furthermore, with recent advancements in generative AI and machine learning, we may be able to lower the burden on content authors by identifying ways to automate different aspects of the content generation and evaluation process, but care must be taken in maintaining human oversight and agency. To unlock the three Cs from the UN's call to action, the next generation of digital learning tools should embody four key themes that can accelerate innovation (Figure 1.2 shows how these themes connect to individual chapters):

1. **Collaboration:** Tools must support collaboration among diverse stakeholders, including instructors, learning scientists, experimental researchers, and data scientists. Given the growing role of generative AI in educational content creation, these tools should also facilitate effective human–AI collaboration, enabling content authors to work productively with AI systems.

2. **Curation:** As stakeholders (and AI) collaboratively explore alternative designs, they need affordances[5] for curating effective formulations into idea pools that future stakeholders can draw from as they continue to improve and adapt content. These idea pools need to be flexible and easy to maintain, so that stakeholders can add versions as new ideas arise.

3. **Comparison:** To curate effective formulations, content authors need ways to compare the relative merits of alternatives using a range of approaches. This could include reflecting on qualitative differences between designs, or running online experiments to gather data on which variations actually improve learner outcomes.

4. **Control:** Finally, tools must help content authors maintain oversight and agency over the content delivered to learners—especially as processes become increasingly automated through emerging technologies such as generative AI or reinforcement learning. Instructors must be able to decide how content is shaped and which versions learners receive.

[5]Here, in the context of the AdapComp framework introduced in Chapter 3, 'affordances' refers to the system-level capabilities and structures that enable stakeholders to curate, maintain, and evolve idea pools effectively. This usage aligns more closely with Gibson's (1979) concept of affordances as action possibilities inherent in an environment [103], rather than Norman's (1988) emphasis on perceived affordances in interface design [242]. Later chapters where I introduce ABScribe and PromptHive align more closely to Norman's characterization.

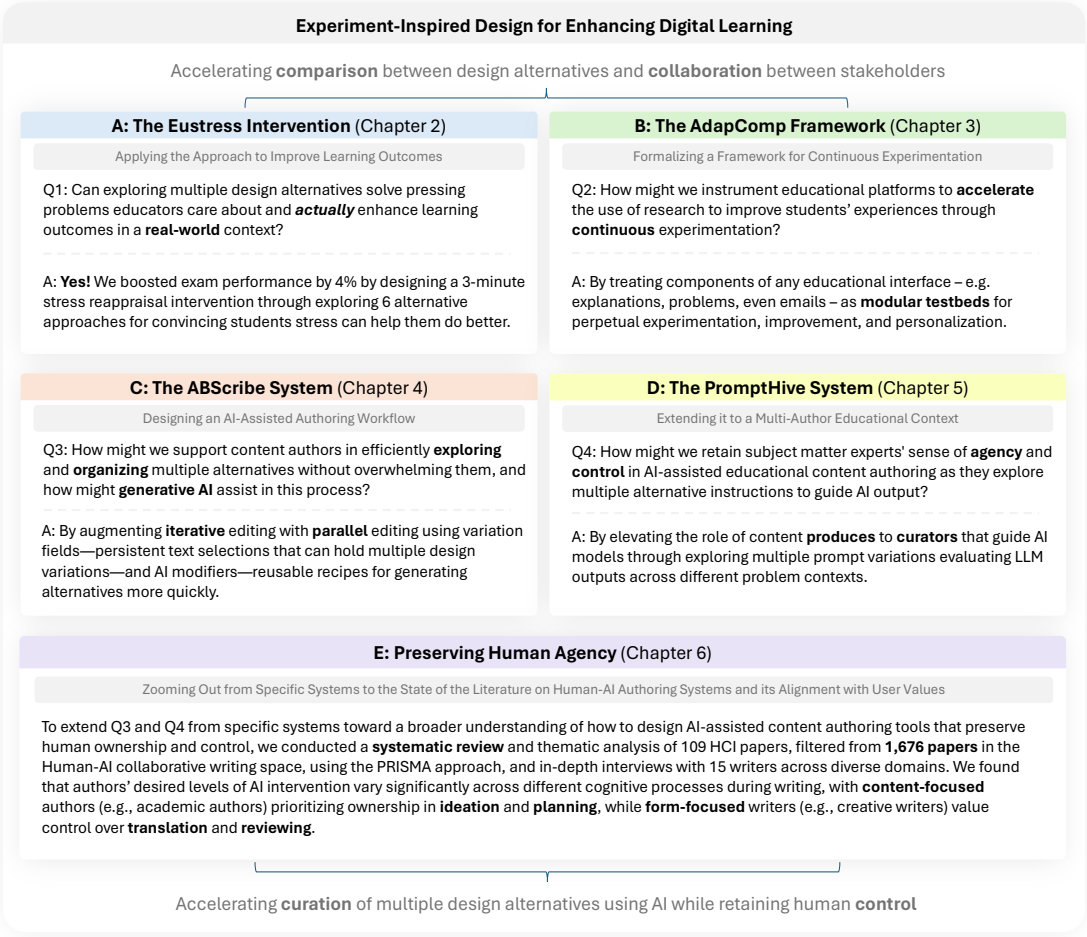


Figure 1.2: Overview of the five projects featured in this dissertation and how they correspond to the four primary research questions and themes for accelerating innovation in digital learning.

[Definition (CR4b): Added a section defining "experiment-inspired design" and its key features.]

1.2 Defining Experiment-Inspired Design

To embody the four themes—collaboration, curation, comparison, and control—within a cohesive methodological approach for enhancing digital learning, this dissertation introduces the concept of *experiment-inspired design*. This approach emphasizes close collaboration between multiple stakeholders in education—instructors, students, researchers, instructional designers, learning engineers, and subject matter experts—to continually improve learning materials by curating an ever-growing collection of alternative learning resources that have been validated through comparison with existing versions. It also emphasizes retaining human control even as we accelerate the generation and evaluation of materials with emerging technologies like generative AI and reinforcement learning. Experiment-inspired design is *inspired* by the logic of traditional experiments involving randomized controlled trials, where carefully constructed alternative conditions are systematically compared.

Akin to how a scientist introduces a new treatment to evaluate its efficacy compared to a baseline condition, an instructor engaging in experiment-inspired design can introduce content variations that encapsulate new ideas and compare them against existing materials to evaluate their relative efficacy. Akin to how the scientific method is theory-agnostic, experiment-inspired design is *pedagogical-framework-agnostic*: it does not prescribe any specific instructional theory. Instead, it enables stakeholders to explore and compare diverse pedagogical strategies—regardless of their theoretical allegiance—within authentic learning environments. Because of this emphasis on validating ideas within authentic learning environments, experiment-inspired design cannot rely solely on the traditional experimentation paradigm. It must be done within the messiness of the real world, not the sterile confines of a lab. And while some aspects of learning can be quantified—such as marks on an exam—others like enjoyment and engagement with learning materials, or how clear an explanation is, often require qualitative judgment. Therefore, experiment-inspired design places equal emphasis on both quantitative and qualitative approaches for evaluating alternative designs. We can therefore characterize experiment-inspired design using the following four key features:

1. **Exploration over confirmation:** Instead of trying to accept or reject some pre-defined hypothesis (as might be the case in a traditional experiment), the goal of experiment-inspired design is to explore a design space through parallel prototyping—generating and evaluating multiple alternative versions of digital learning materials.

2. **Qualitative as well as quantitative:** Experiment-inspired design values both numerical indicators of learning (e.g., learning gain, marks, click-through rates) as well as qualitative evaluations (e.g., semi-structured interviews, qualitative reflections on alternatives from experts, student reflections on enjoyment) as valid sources of insight.

3. **Continuous, not one-off:** Experiment-inspired design treats experimentation as a continuous process as opposed to an isolated event. Alternative materials are continually improved through rapid iteration, with each comparison fueling new ideas for further refinement.

4. **Field-based, not lab-based:** Such rapid iteration needs to happen within authentic learning environments embedded within actual digital learning platforms. Stakeholders need to work within the constraints of the real world, and factor in how those constraints shape user behaviours and outcomes.

[Introduction (CR1a): Added a section on the overarching theories that this work draws from and contributes to, discussing how it relates to different pedagogical and teaching frameworks.]

1.3 Theoretical Foundations

This dissertation is grounded in and contributes to several interconnected pedagogical, research, and design traditions. Chief among these is the lineage of design-based research (DBR) [36], a methodological approach from the learning sciences that involves running "design experiments" in field settings to answer questions that laboratory-based investigations cannot adequately address [17]. Like DBR, experiment-inspired design focuses on iterative cycles of continuous improvement, co-developing solutions (or "interventions") through collaboration among multiple stakeholders. Like

DBR, experiment-inspired design emphasizes artifact design over hypothesis testing and responding to situated educational challenges. Experiment-inspired design shares the primary goal of DBR in advancing theory through the design and analysis of interventions [9]. At the same time, experiment-inspired design puts greater emphasis on improving digital learning resources through the systematic comparison of alternatives. DBR in practice is typically qualitative and richly contextual, whereas experiment-inspired design promotes a mix of quantitative (e.g., A/B testing) and qualitative comparisons (e.g., reflecting on qualitative differences between variants) in parallel. DBR is less formalized–often messy and narrative–whereas experiment-inspired design supports more structured and modular improvement of course components, with a focus on parallel prototyping and continuous iteration.

Second, it draws ideas from learning engineering [109, 318], and shares its emphasis on using a human-centered design [340] approach coupled with analyzing rich datasets to enhance digital learning. It also reflects the interdisciplinary outlook of learning engineers, who often combine principles from computer science, data science, and learning science and apply those principles to real educational settings [341].

Third, this work is informed by ideas from online multivariate A/B testing [180], which, like experiment-inspired design, involves a structured comparison of alternatives in digital settings. Like A/B testing, experiment-inspired design relies on empirical evidence to evaluate efficacy and requires systems for deploying multiple variants, collecting data, and analyzing results—whether the content is on a webpage or instructional materials in a course. Both use the concept of comparing new alternatives against a control or baseline condition. However, experiment-inspired design differentiates itself from online A/B tests in its focus on enhancing digital learning rather than general product improvement (akin to the distinction between learnersourcing and crowdsourcing). It also differs in terms of evaluation style: A/B testing is primarily quantitative, whereas experiment-inspired design integrates qualitative feedback and reflection on comparisons. The approaches also differ in their disciplinary roots—A/B testing originates in product engineering, marketing, and web UX, whereas experiment-inspired design draws from HCI, learning sciences, and design theory.

Finally, this work draws from HCI and traditional design practices that explore design spaces by creating and comparing multiple alternatives—both to avoid fixation on a single idea [151, 108] and promote more objective appraisal through critical comparison between alternatives [345, 42]. Experiment-inspired design adopts parallel exploration for similar reasons but shifts the focus from low-fidelity sketching to high-fidelity, in-field testing. It emphasizes implementation in authentic learning environments and is supported by the technologies developed in this dissertation for creating and evaluating content variations. Its dual emphasis on retaining the rigor of controlled experimentation while incorporating qualitative assessments to support richer, more ecologically valid experiences aligns with recent methodological characterizations in HCI, such as the idea of a Comparative Structured Observation (CSO) [213]. CSO highlights the value of gathering comparative reflections on design variants to help researchers assess and refine their concepts—an argument this dissertation extends to the context of digital learning materials.

[Introduction (CR1b, CR1c): Added a section stating the overarching research question and scientific contribution, emphasizing how the different chapters address various facets of the challenges involved in authoring and evaluating multiple design alternatives to enhance digital learning. Included a table to better highlight each chapter's contribution and the common thread connecting

them.]

1.4 Overarching Research Question and Contribution

The overarching research question this dissertation seeks to answer is:

> *How can digital learning materials be continuously improved through the systematic exploration and evaluation of multiple design alternatives?*

In this thesis, I show that facilitating the parallel exploration and evaluation of multiple design alternatives through experiment-inspired design can accelerate the enhancement of digital learning materials and lead to demonstrably better learning outcomes. I first demonstrate the impact of this approach through the design of the **Eustress Intervention** (Chapter 2), where I apply experiment-inspired design to tackle the problem of exam stress. Then, I develop a collection of human-centered tools that support both the rapid exploration of content variations in **ABScribe** (Chapter 4)and the evaluation of those variations using manual and automated means in **AdapComp** (Chapter 3). I also show how AI can be integrated into the content authoring process without ceding human control, through systems like **PromptHive** (Chapter 5) and a broader investigation into human agency in AI-assisted workflows (Chapter 6).

Together, these projects address the two interconnected challenges in digital learning mentioned earlier: (1) the lack of *authoring workflows* that support the parallel creation and management of multiple content variations, and (2) the lack of *deployment mechanisms* to systematically deliver and evaluate those variations in authentic learning environments.

Each chapter of the dissertation contributes to tackling different facets of these challenges, as summarized in Table 1.1. **ABScribe** (Chapter 4) and **PromptHive** (Chapter 5) provide concrete authoring workflows for managing multiple content variations. In ABScribe, I address this challenge in the general context of writing, introducing an interface that explicitly supports the exploration and organization of variations—accelerated through the use of generative AI. PromptHive extends these ideas to a collaborative setting involving multiple domain experts, shifting the focus from writing broadly to designing educational content with measurable impacts on learning outcomes.

The **AdapComp** framework (Chapter 3) tackles the second challenge: deploying and evaluating content variations systematically. Through a flexible API-based architecture, AdapComp integrates with real-world digital learning platforms such as Canvas and OLI, and provides infrastructure for adaptive experimentation, reinforcement learning, and version control over instructional content.

Finally, because automation via AI/ML is central to many of the systems developed in this thesis, I dedicate **Chapter 6** to a systematic review of human–AI collaborative writing. This review synthesizes design strategies for retaining human agency and control in AI-assisted content authoring workflows.

1.5 Thesis Statement: Using Parallel Exploration

To put it simply, the overarching research objective of this dissertation is to develop better approaches and tools for *continuously enhancing digital learning materials*. My thesis is that:

#	Chapter	Focus & Challenge	Key Contribution
2	**The Eustress Intervention**	Demonstrating impact of experiment-inspired design by tackling exam stress as a real-world challenge *(Authoring + Deployment)*	Shows real-world learning gains achieved through applying the approach
3	**The AdapComp Framework**	Establishing infrastructure for continuous experimentation-driven improvement *(Deployment)*	Adaptive experimentation framework integrated with real LMS platforms using APIs
4	**The ABScribe Interface**	Designing a Human-AI Authoring Workflow for Parallel Exploration *(Authoring: Individual)*	Human-AI interface for parallel revision; lowers cognitive load during exploration
5	**The PromptHive System**	Involving Subject Matter Experts in AI-Asssisted Authoring *(Authoring: Collaborative)*	Extends ABScribe for multi-author workflows focused on improving educational content
6	**Preserving Human Agency**	Retaining Human Agency in AI-Assisted Authoring *(Authoring: Individual & Collaborative)*	Systematic review + interviews provide design strategies for preserving authorial control with AI

Table 1.1: How each chapter contributes to solving the two core challenges of experiment-inspired design: (1) the lack of *authoring workflows* that support the parallel creation and management of multiple content variations, and (2) the lack of *deployment mechanisms* to systematically deliver and evaluate those variations in authentic learning environments.

> *Facilitating parallel exploration and evaluation of multiple design alternatives can accelerate enhancement of digital learning materials, leading to better learning outcomes.*

To frame and guide this investigation, we[6] break the core research question into four primary sub-research questions that are interlinked with the themes outlined earlier on accelerating innovation through experiment-inspired design. The first question centers around whether this approach can actually impact learning outcomes in a real-world context. The other three are framed as "How-Might-We" questions[7], a design thinking method that facilitates problem-solving and fosters empathy by encouraging "designers and researchers to step into users' shoes, and consider their challenges, needs, and desires [96]."

1. Can exploring multiple design alternatives help solve pressing problems that instructors, learners, and other stakeholders care about and enhance learning outcomes in real-world contexts?

2. How might we instrument educational platforms to accelerate the use of research to improve students' experiences through continuous experimentation?

3. How might we support content authors in efficiently exploring and organizing multiple alternatives without overwhelming them, and how might generative AI assist in this process?

4. How might we retain domain experts' sense of agency and control in AI-assisted educational content authoring as they explore multiple alternative instructions to guide AI output?

[6] I use the term *we* to emphasize the collaborative nature of my research. This includes not only the invaluable contributions from my co-authors but also the instructors and students whose participation has been crucial in shaping and answering these research questions.

[7] "The 'how' part assumes there are solutions out there — it provides creative confidence. . . 'Might' says we can put ideas out there that might work or might not — either way, it's OK. And the 'we' part says we're going to do it together and build on each other's ideas." — *Tim Brown, IDEO Chairman and former CEO* [2]

1.6 Thesis Overview and Research Outline

To address these research questions, we took the following steps:

1. To answer Question 1, we designed the *Eustress Intervention* (Chapter 2), a collection of digital learning resources for tackling exam stress—a pressing and pervasive problem that many learners and instructors care about. We did this by exploring six alternative design factors, and evaluated our design through 20 user interviews and an online field experiment with over 1,000 students in a real-world course. Our results highlight the power of this approach, as the resources we designed significantly boosted exam performance by 4% ($p = 0.003$), even though learners only had to spend an average of 3 minutes reading the resources.

   [Motivation (CR3a): Clarified that the purpose of the first study on exam stress is not to emphasize the topic itself, but to demonstrate a methodological approach (experiment-inspired design)—developing and evaluating multiple content variations in parallel. This study serves as a proof-of-concept for the broader challenge addressed in the thesis.] Crucially, while this study focuses on exam stress reappraisal, its purpose within the dissertation is not to emphasize the importance of that specific topic, but to illustrate how educational interventions for real-world student challenges can be developed through the parallel creation and evaluation of multiple content variations. Experiment-inspired design is *pedagogical-framework-agnostic* (see Section 1.2). Thus, this chapter serves as a concrete proof-of-concept for the broader methodological contribution of the dissertation. Rather than committing to a single version from the outset, we systematically explored a design space of six reinforcing components, refined them through qualitative feedback, and evaluated their combined impact using a quantitative experiment in a real course setting. This structured exploration mirrors the kind of content experimentation that the rest of the dissertation aims to scale through dedicated authoring tools and infrastructure.

2. To answer Question 2, we formalized the *AdapComp Framework* (Chapter 3), an open-source web service for enhancing digital learning resources through a wide range of manual and automated approaches using reinforcement learning. This framework transforms digital learning resources into modular testbeds where instructors, experimental researchers, and data mining/machine learning researchers can engage in perpetual cycles of experimentation, improvement, and personalization.

3. To answer Question 3, we designed the *ABScribe Interface* (Chapter 4), an authoring interface that augments conventional sequential text editing approaches with a novel interface for parallel editing. ABScribe supports users in organizing and exploring many writing variations in an interface that still feels familiar to them, and leverages AI to help accelerate the text revision process while keeping workload low. Results from an evaluation study with 12 content authors show that ABScribe significantly reduces task workload ($d = 1.20$, $p < 0.001$) and enhances user perceptions of the revision process ($d = 2.41$, $p < 0.001$) compared to a popular baseline workflow.

   [Motivation (CR3b): Clarified that ABScribe is designed to support the text revision process more broadly—not just for students, but also for instructors and other writers. Explained how its features integrate with existing rich-text editing environments and enable instructors

to iteratively refine instructional content, directly supporting the thesis's overarching goal of enhancing learning through parallel content exploration.] This interface is geared toward supporting the text revision process more broadly and is designed for writers in general—including instructors, journalists, fiction and non-fiction authors, researchers, and students. Its interface features—such as Variation Components, Hover Buttons, the Variation Accordion, AI Buttons, and AI Insert—are designed to integrate seamlessly into existing rich-text editing environments, including text-entry panels within course management systems (e.g., D2L Brightspace, Moodle, Google Classroom) and general-purpose editors such as Google Docs. The goal is to offer a familiar editing experience while enabling users to move beyond producing and publishing a *single* version of content. For example, instructors can use ABScribe to iteratively improve their explanations and explore alternative ways of communicating ideas to students—contributing directly to the broader research goal of enhancing the learning experience through parallel exploration. Like the interface elements, the underlying foundation models are also general-purpose, allowing the system to be adapted to a variety of learning contexts.

4. To answer Question 4, we developed the *PromptHive System* (Chapter 5), a collaborative system that extends some of the ideas from ABScribe for prompt authoring, designed to better connect domain knowledge with prompt engineering through features that encourage rapid iteration on prompt variations. We conducted an evaluation study with ten subject matter experts in math, as well as a learning gain study with 358 learners. Our results validate the tool's usability, enabling non-AI experts to craft prompts that generate content comparable to human-authored materials, while reducing perceived cognitive load by half and shortening the authoring process from several months to just a few hours.

   [Motivation (CR3c): Clarified how PromptHive can be generalized to support instructional scenarios where instructors author, revise, and curate AI-generated educational content, and how the need for such out-of-class authoring is becoming increasingly apparent in light of recent shifts in instructional practice (particularly the rise of unsupervised blended learning spaces.] Recent shifts in instructional practice—particularly the rise of unsupervised, blended learning spaces [188] such as Canvas-based homework [44], formative assessments, and systems like PCRS [269]—have introduced new challenges for instructors. These environments are increasingly AI-supported and instructor-authored, yet fall outside traditional classroom evaluation and feedback workflows. PromptHive addresses this gap by empowering instructors, alongside other subject matter experts, to take an active role in designing, testing, and refining instructional prompts—for example, hints or explanations—in these out-of-class digital spaces. While our case study applies PromptHive in the context of an adaptive tutoring system (OATutor), its core interface and workflow are generalizable to instructional scenarios where instructors author, revise, and curate AI-generated educational content. By enabling structured and collaborative experimentation with alternative prompt versions using ideas from ABScribe—while maintaining expert oversight, supporting rapid iteration, and offering curation tools—PromptHive directly operationalizes the four principles of experiment-inspired design: collaboration, curation, comparison, and control.

5. To move from specific systems toward a broader picture of how to thoughtfully automate

content authoring workflows using generative AI while preserving human agency (Chapter 6), we go deeper on Question 3 and 4 by conducting a systematic review of 109 HCI papers, distilled from 1,676 recent studies in the Human-AI collaborative authoring space using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. We complement this review with user perspectives from in-depth interviews with 15 writers across diverse domains. Our findings reveal that writers' desired levels of AI intervention vary significantly across these processes, with content-focused writers (e.g., academic authors) prioritizing ownership in *ideation* and *planning* process, while form-focused writers (e.g., creative writers) value control over *translation* and *reviewing*. We also identify four overarching design strategies, such as adaptable AI intervention and explicit control mechanisms, that align AI support with writers' values tied to originality, agency, and ownership.

Chapter 2

The Eustress Intervention

> 80-90% of education interventions don't work when tested rigorously.
>
> *Stuart Buck [298]*

Research Context: This chapter describes an intervention we designed and deployed in a large first-year CS course right before the peak of the COVID-19 pandemic. The goal was to see whether we could use our approach of exploring multiple design alternatives in parallel to help students manage exam stress, a widespread problem that was further exacerbated during that time due to the uncertainties surrounding the unforeseen shift to online learning. This intervention servers as an example that addresses our first of four primary research questions:

> *Q1: Can exploring multiple design alternatives help solve pressing problems that instructors, learners, and other stakeholders care about and enhance learning outcomes in real-world contexts?*

The answer is, fortunately, a resounding yes. We were able to take counterintuitive insights from decades of psychology research on stress reappraisal, extend them to a new real-world context within months, and developed a powerful 3-minute intervention that boosted exam performance by 4%.

Publication:

Mohi Reza, Angela Zavaleta Bernuy, Emmy Liu, Tong Li, Zhongyuan Liang, Calista K Barber, and Joseph Jay Williams. 2023. *Exam Eustress: Designing Brief Online Interventions for Helping Students Identify Positive Aspects of Stress.* In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 439, 1–13. https://doi.org/10.1145/3544548.3581368

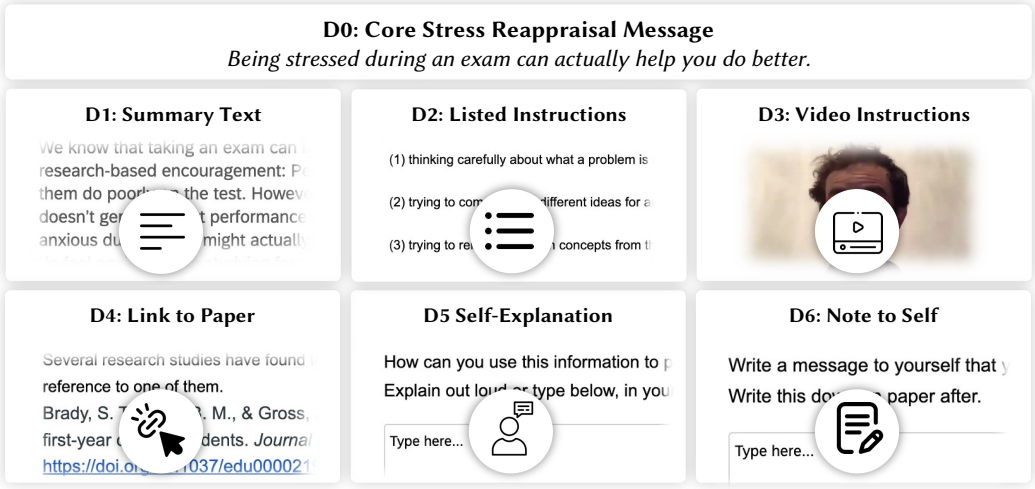


Figure 2.1: Our design space for exam eustress interventions consists of a core stress reappraisal message (D0) reinforced by 6 design factors: (i) D1 offers explanatory context for reappraisal in paragraph form, (ii) D2 gives explicit suggestions for what to do during exams, why stress could help, and how to use this information, (iii) D3 includes a talking-head video from an instructor explaining the explicit suggestions (iv) D4 provides a citation and link to a research paper, (v) D5 prompts students to self-explain the concept by typing or voice, (vi) D6 prompts students to write a note that they could revisit before the exam.

Abstract: Stress reappraisal interventions try to shift students' negative perceptions towards *eustress*, stress that can be beneficial, and help them perform better. However, it is less clear how to present them to users as online interventions that are brief, voluntary, and scale well in real-world contexts. We explore the design of online *exam eustress* interventions by generating six design factors (D1-6) that reinforce a core reappraisal message (D0), and evaluate them through: (i) user interviews ($N = 20$) revealing six findings (F1-6) on the importance of elaboration, layout, modality, and source of intervention content; (ii) a field experiment ($N = 1283$) showing a significant positive effect on exam scores ($p = 0.003$). Subgroup analyses indicate a significant effect for first-year but not for upper-year students, and no detectable gender differences. Our work offers insight into how students interact with online mindset interventions and design considerations for incorporating them into large courses.

2.1 Introduction

Stress has become increasingly prevalent in higher education, and exams are a key stressor[289]. While COVID-19 has further exacerbated the stress levels of students [326], it has also accustomed them to using digital learning tools [74], creating opportunities for instructors to leverage online interventions. These interventions can be used to help students not only reduce stress but also to find a way to use stress as a resource for handling challenging situations such as exams. Students often interpret normal physiological reactions to stress as wholly harmful. These negative perceptions about stress and its effects on test performance can lead to increased *distress*.

Introducing students to the concept of *eustress*, i.e. a positive psychological response to a stressor,

has the potential to help them reappraise stress as not necessarily something to be eliminated but something to be embraced as a signal that the body is trying to help them focus and perform better. There is increasing evidence that this kind of messaging can benefit students in exam situations, such as when they are required to read some materials [149] or when they receive information right before the exam [32]. In this paper, we explore how to provide online activities that support students by helping them change how they think about stress's benefits.

For this exploration to be constructive, we must be aware of the many design constraints and considerations in developing online activities for students that communicate reappraisal messaging in an actionable way. In this paper, we explore the design space for activities that are **simple** (single webpage), **brief** (takes less than 5 minutes), **scalable** (requires no instructor feedback), and equip students with the ability to apply this information to their lives through self-reflecting on how to leverage eustress during exams. We think these particular elements are essential because they make online interventions feasible and impactful in real-world contexts, where students may have limited attention, and instructors are often overloaded with limited capacity for one-on-one interactions, despite wanting to help students.

Within this challenging context, we explored the design of multi-component online interventions by varying several complementary factors to reinforce engagement with a core stress reappraisal message – being stressed during an exam can help students do better. We varied the factors in terms of three design dimensions: (i) the *amount of information* and explanatory text (balancing the trade-off of limited attention and visual clutter against the degree of elaboration on how stress can be helpful), (ii) the *modality of presentation* (text versus video), and (iii) the potential value of *reflection prompts* for students to think aloud and type notes on how their future selves could use the information during their next exam.

We evaluated the range of design components through semi-structured interviews with 20 students and a randomized field experiment deployed to 1283 students in a programming course. This provided insights into the relative importance of different factors and their impact on different students in different contexts, such as when text versus video presentation might be effective, what kinds of additional information are compelling versus redundant, and what kinds of interface prompts for students are more or less impactful in terms of helping students retain and utilize the reappraisal messaging. Although the activity was delivered via a brief and optional online intervention that took an average of under 3 minutes, it positively impacted student performance on a test, with a significant effect on first-year students but not upper-year students. Our findings provide insight into how instructors and others can design online mindset interventions based on contextual features of the interface and student characteristics. We also offer design directions for a range of potential future work in the better digital delivery of mindset interventions. The main contributions of this work are:

1. An exploration of the design space for online exam eustress interventions that are simple, brief, and scalable, focusing particularly on how students interact with different presentation modalities (text, video, or both), levels of elaboration (explanation, instruction, or paper citation), and reflection prompts to learn how to identify beneficial aspects of stress during exams. Our methodology can also be applied to design other scalable stress-management interventions.

2. An evaluation of our multi-component design using semi-structured interviews with 20 stu-

dents, as well as a large-scale online field experiment with over 1000 students providing evidence on the positive impact of our intervention on exam performance and insight into how students interact with various components described in figure 5.1.

2.2 Related Work

2.2.1 Psychological Research into Practical Technology-Mediated Interventions

There is accumulating evidence that, in specific contexts, stress reappraisal strategies can be effective for managing feelings of worry about anxiety [146, 133] and enhancing performance in situations of acute stress [146], such as exams [148, 147, 32, 206]. Past research has shown examples where videos [64], emails [32], and text instructions can be better than no reappraisal. However, it remains unclear how to combine them into brief online interventions because there has been little direct comparison between such presentation modalities. We explore the design space of different modalities, explicitly comparing video, text, and a combination of video and text. Given how important it is to preserve student attention and to empower instructors to help more students without becoming overloaded, we focus our exploration on voluntary, brief and scalable intervention combinations. While it is particularly challenging to design interface components that meet all of these three criteria, we think they are essential because interventions that meet them have the potential to translate psychological research into *practical* technology-mediated interventions that can positively impact thousands of students.

2.2.2 HCI Research on Prompting People to Change their Beliefs and Attitudes

We contribute to the growing body of existing work within HCI that focuses on leveraging online technology to promote positive belief change at scale by designing systems that apply various behavioural science approaches to real-world problems such as mental health treatment using online chat tools [245, 18] and conversational coaches [216], engaging users to reflect on physical activity [175], cognitive behavioural therapy (CBT) [89] based apps to reduce depression and anxiety [114, 30], and just-in-time (JIT) based interventions [121] to reduce digital workplace stress [135]. In this paper, we present a design exploration of how different content modalities can be used to communicate stress-reappraisal messaging to students in large classrooms effectively. We give qualitative insight into how students interact with different modalities to alter their stress mindset and quantitative evidence showing that our designs improve exam performance in a large-scale field experiment.

2.2.3 Shifting from Designing *Distress* Interventions to *Eustress* Interventions

Current HCI research in the online intervention space has focused primarily on applying *reduction-based* coping strategies to manage stress instead of trying to help users embrace positive *eustress*[189]. We believe this reflects the dominant historical orientation of past research toward negative aspects of

stress. For example, previous studies [67, 48] have demonstrated that prolonged stress lowered overall academic performance [67, 48], as indicated by low exam scores and low overall grade point average. These studies suggest that test anxiety, a common source of stress, is a cause of poor academic performance as it negatively affects critical factors that affect the learning process, such as sleep and biological systems that mediate the body's responses to stress. However, building upon pioneering work from the 1980s [95], a new class of studies has hinted at the multiple potential benefits stress can have on academic performance when used methodically [83, 21, 331, 205]. Compared to most existing HCI contributions in the stress-management space, one contrasting feature of our intervention is that it focuses on reappraisal rather than stress reduction.

Our decision to focus on the positive aspects of stress is informed by previous studies on emotional regulation that have compared the use of reappraisals for stress management with other strategies such as suppression [133, 157], self-distraction [84, 154], and acceptance [133, 107, 348]. These studies suggest that reappraisal is a more effective strategy than suppression and acceptance for moderating physiological arousal and the subjective feeling of stress [133]. The reappraisal strategy can also help reduce the negative emotional experience of stress [84, 206]. Both reappraisal and self-distraction may be effective for attenuating emotional reactions [154]; however, we choose reappraisal for our intervention as we aim to measure its effectiveness in an exam setting.

2.2.4 Situating Design Factors in Prior Work

To generate our six design factors, we draw guidance from research on multimodal information presentation in HCI [46, 349], effective communication strategies from market research [219], instructional design [181, 155], multimedia learning [218] and reflective learning [344]. Convincing and concise reporting often requires incorporating data from multiple sources, methods, and modalities [219, 46, 218]. To leverage this approach, we layer design factors on top of each other to reinforce a core reappraisal message. Literature on instructional design and multimedia learning informs our choice to test different content layouts (e.g. paragraphs and bullet points) and mediums (e.g. text and video). Prior work has shown the potential instructional benefits of spoken words in videos [155], how bulleted lists may improve information retention [142, 38], and how mixing modalities could improve learning by offering learners parallel opportunities for information processing [388]. One of the biggest design challenges is how to help learners retain the reappraisal information and potentially change their behaviour on the upcoming exam. To address this, we turned to prior work on reflective learning via writing and voice, which has shown the success of reflection prompts in guiding future behaviour. We incorporate these insights in the final part of our intervention by designing two reflection prompts for assisting recall [246] and behaviour change [344, 90].

2.3 The Design Space for Online Exam Eustress Interventions

2.3.1 Design Constraints and Considerations

We want to use online interventions to impact students' behaviour by helping them rethink exam stress as being useful rather than detrimental. We consider the constraints posed by brief online

webpage interfaces. On one hand, such interfaces are ideal for technology-mediated online intervention because students can easily access them via their computer or smartphone browsers. However, for an intervention embedded within such a setting to be effective, it needs to be (i) *voluntary*, because students don't *have* to do them, and (ii) *scalable*, so it can be sent to many students without requiring instructor intervention. These two constraints are considerably challenging to meet due to the counter-intuitive nature of the reappraisal message. One could imagine that reappraisal messaging is best delivered in person, by someone talking to the students, explaining the idea to them, sharing stories, and asking questions.

To investigate the design of components of a brief, simple, scalable, online intervention interface that can be self-administered by students, we considered content modalities and information that target three elements of users' cognition and behaviour:

1. Information presentation that communicates the message effectively to students.
2. Information content that engages them in deeper processing to understand the information.
3. Prompts them to consider specific actions so they would be more likely to remember the ideas in the future.

A key guiding consideration is that students' attention in online environments may be incredibly limited and that instructors are often overloaded and unable to offer one-on-one in-person interactions with students at scale. Therefore we asked what the considerations are in deciding what information and activities merit being included to convey a reappraisal message that is effective for different students in varying contexts, which accounts for the trade-off between limited student attention and offers a thorough explanation, and results in students internalizing the message.

2.3.2 The Six Design Factors

In section 2.2.4, we discussed how our intervention design factors drew insights from prior work on multimodal information presentation in HCI [46, 349], effective communication strategies drawn from market research [219], instructional design [181, 155], multimedia learning [218] and reflective learning [344]. In this section, we revisit some of those insights and describe the specifics of our design and our rationale behind each factor.

D0: Core Stress Reappraisal Message. All students received the following core message:

> *"Being stressed during an exam can actually help you do better."*

We suspected that this brief message alone would not be sufficient in convincing students that stress can be helpful, which was later confirmed in our user interviews (see F1 in section 2.4.4). An effective communication strategy from market research involves the integration of multiple sources and methods into well-synthesized content [219]. We adopted this strategy and explored ways to reinforce D0 using six design factors (D1-6).

D1: Explanatory Elaboration Providing Research-based Rationale and Encouragement. We provided additional explanatory context by laying out the specific logic of reappraisal to understand how important this information was to students, and which aspects of this elaboration were useful in different contexts. Furthermore, we wanted to evaluate whether framing the stress reappraisal concept as being *research-based* would convince students, in alignment with science

communication literature on conveying accurate scientific information through persuasive scientific narratives [68].

The text that could be included or not was:

> *"We know that taking an exam can be a stressful experience, and so we wanted to provide a note of research-based encouragement: People think that feeling anxious while taking a test will make them do poorly on the test. However, recent research suggests that increased levels of stress doesn't generally hurt performance on tests and can even help performance. People who feel anxious during a test might actually do better. This means that you shouldn't feel concerned if you do feel anxious while studying for or taking the upcoming exam. If you find yourself feeling anxious, simply remind yourself that your stress and its higher energy could be helping you do well."*

D2: Explicit Suggestions for What to Think During Exams, Why Stress Could Help, and How to Use the Information. While offering explanatory elaboration may help convince students, we wondered if they needed explicit suggestions on what to do. Our rationale was to evaluate whether it was helpful to guide students through directed prompts [73, 335] containing specific instructions and explanations on how to reappraise stress. We also wanted to see if this information was redundant or even potentially unhelpful as it could reduce student attention and dilute the impact.

> *"During your exam, try to remember that feeling stressed might actually help you perform better, by making you more alert, and helping you work harder.*
>
> *Try to use the feeling of stress as a cue, to put energy into: (1) thinking carefully about what a problem is asking you, (2) trying to come up with different ideas for answering questions, and (3) trying to remember which concepts from the semester are relevant.*
>
> *How might stress help you do better on an exam? Your brain is recruiting resources to make you pay attention, so that you can have more energy to work hard and think deeply.*
>
> *If you find yourself feeling stressed during the exam, remind yourself that this is normal, and not necessarily bad – it may even be helping you do better than if you weren't stressed."*

In light of prior studies showing the benefits of spoken words in instructional videos [155] and the potential to improve information retention and learning through mixing modalities [246], we wondered whether varying how the explicit suggestions were presented to learners, either as text or video, would be helpful in our context.

D3: Presentation by Instructional Video. We developed an instructional talking-head video where a faculty member researcher explained the concept in an enthusiastic manner in order to explore how students' preferences for the modality varied and if there was value in both. For simplicity of presentation, we chose to do it for just the explicit suggestions from D3.

D4: Validation by Explicit Citation of a Source Paper. Providing explicit evidence that research studies support stress reappraisal, versus simply giving the explanatory elaboration of why this could be useful.

> *"Several research studies have found that sharing messages like this can help people do better on exams – here is the reference to one of them.*
>
> *Brady, S. T., Hard, B. M., & Gross, J. J. (2018). Reappraising test anxiety increases academic performance of first-year college students. Journal of Educational Psychology, 110(3), 395–406. https://doi.org/10.1037/edu0000219"*

Finally, to tackle the considerable challenge of helping students retain and apply the intervention information during their exam, we turned to the literature on behaviour change through reflective learning [344, 90] and designed two prompts that encouraged students to consider how to apply this information moving forward.

D5: Prompt to Reflect on How to Use the Information During the Exam. Self-monitoring or generic prompts encourage students to reflect and produce more coherent ideas than directed prompts [73]. There have also been studies showing the benefits of self-explaining through voice or writing [246]. Therefore, we designed a self-monitoring prompt where students were told they could talk out loud, type, or both.

> *"How can you use this information to perform well on your exam? Explain out loud or type below, in your own words"*

Our goal was to prompt students to deepen their understanding of the information by being active instead of passive.

D6: Prompts to Type out a Message to Look at Before the Exam and Write It on Paper. To maximize information retention and chances of behaviour change during the exam, we asked students to write a message that they could revisit before the exam.

> *"Write a message to yourself that you can look at right before the exam, as a reminder of how to use this information. Write this down on paper after."*

We wanted to help students visualize what they could say to themselves before the exam to further increase the chances they remember the reappraisal message. It was less important whether the students actually wrote it down.

These factors explored a complex set of design components, but suggest many directions for other components that could be explored in future research. We discuss some of these directions in section 2.6.3. Each factor was presented in the same order as their names, i.e. D0 to D6. We chose this order because it illustrates the design principles from prior work as a logical structure, i.e. having the core reappraisal message at the very beginning, augmented by some explanatory text specifying the underlying logic (or not), some concrete steps for using the knowledge (or not), a video on those steps (or not), a paper citation for added credibility (or not), and finally, the two reflection prompts (or not). To evaluate our design, we conducted two studies, as discussed in sections 2.4 and 2.5.

2.4 Study 1: User Perspectives on Intervention Components

The first study was an exploration of user perspectives to characterize the impact and trade-offs of including particular components of the design. We conducted interviews with students where we showed all the different components of the six design factors of our online intervention interface. We

asked users to reflect on each component and share what they thought or felt as they read it, to better understand the impact and trade-offs of including a particular component of the design.

2.4.1 Participants

Our participants consisted of 20 students (14 women and 6 men) aged 18 and above, from a large and diverse introductory programming course. We recruited them through a call for participation via email to students who previously expressed interest in research activities. To gather diverse perspectives, we recruited students from different years and disciplines. In terms of their year in the program, ten participants (P2-6, P8-12) were starting their second year, six were going into their 3rd year (P1, 7, 13, 14, 17, 19), three were in their 4th year (P8, 16, 18), and one just graduated (P15). Participants were studying a range of subjects including Accounting, Actuarial Science, Biology, Computer Science, Economics, Mathematics, Physics, Physiology, and Statistics. Their diverse academic backgrounds sparked rich, wide-ranging conversations about exam stress that were not tied to viewpoints from a specific discipline or year in the program.

2.4.2 Procedure

Participants attended the interview using online video conferencing software. They completed a consent form and gave us permission to record the session before starting. Each interview lasted between 45 and 60 minutes, and participants were compensated 15$/hr for their time. The interviewer started with a brief conversation about the participants' past experience with exam stress to understand any pre-existing notions they may have about whether stress is good, bad or both. After this conversation, the participant was asked to share their screen, and to go through every component (D0-6)–first individually on separate webpages, so we could understand how they felt about each component in isolation, and then all on one webpage, so we could understand how they compared components with each other. When going through the components individually, we counterbalanced the order in which participants saw the text (D2) and the video (D3) because we were interested in understanding subjective user impressions immediately after seeing each factor. As the participant went through the activity, the interviewer asked them to think aloud and share any thoughts and feelings as they arose. A silent notetaker was also present during the interviews to write down observations.

2.4.3 Analysis

Our qualitative data consisted of interview transcripts that were generated from the recordings, observation notes, and video recordings. We coded and analyzed this data using reflexive thematic analysis [33] through an inductive lens, drawing from the rich theory on stress mindset and reappraisal as a pre-existing code.

2.4.4 Findings

Our findings (F1-F6) indicate a clear need for reinforcing the core reappraisal message presented in D0 with the six design factors (D1-6). We order them considering their prevalence in our data and our judgement on their importance.

F1: A Short Stress-reappraisal Message on Its Own Is Not Sufficient Enough to Convince Users.

> *"D0 is just a small sentence, I think. It is just conveying a conclusion but we don't know how this conclusion came about. I don't think that it is a good choice to convey this idea on its own." – P2*

Based on comments from several participants (P1-4, 6, 8, 10-12), we found that D0 alone is not sufficient for convincing users that stress can be beneficial to them because user agreement with the brief stress reappraisal message in D0 largely depended on past experience with stress. P1 and P2 noted how D0 was too short and did not offer an explanation of how it reached the conclusion that stress can be helpful during exams. P3 and P4 mentioned that they mostly agreed with D0 because of their past positive experience with stress. F1 is further exemplified through comments from those who did not initially agree with D0, such as P8, 10-12. After expressing disagreement, they described how past exams where they experienced stress did not go well.

We also found that users could potentially misinterpret the message in D0 when presented alone because of how brief and counterintuitive it was. After reading the message, P6 said that they found it a "bit weird because everyone would experience stress during the exam" and "with the way D0 is framed, it's like we should try to eliminate stress entirely". The latter half of the comment clearly shows that the user took away the exact opposite message to the one we were trying to convey, i.e. not to eliminate stress but rather to embrace it as being helpful. Furthermore, users also appeared to have missed the point about timing. The stress reappraisal message referred to stress *during* the exam. However, among those who said they agreed with the message, such as P3, further discussion on why they agreed with it revealed that they were thinking of stress before the exam and how that stress helped them prepare better.

F2: Users Find Content That Is Described as Being Research-based More Convincing.

> "I think stress and anxiety are different, but having read D1...it will be more convincing...because there is research showing that this idea, this conclusion about stress is not something that people just imagined, but did research on to come to the conclusion." – P2

A noteworthy element of the wording in D1 that stood out to participants was that we mentioned how the suggestions contained in the message were *research-based*.

As exemplified by P2's comment above, participants felt convinced by D1 because we mentioned that it was research-based. P2 felt convinced by D1 because it "summarizes research about stress during exams". Similarly, P1 said "knowing that this is research-based. . . I don't think I would feel like it would be a bad thing". They even mentioned, "I think I would be grateful for a little bit of stress when writing my exams". P8 mentioned that their thoughts on anxiety changed after seeing D1 "because research suggests that feeling anxious during your exam could lead to better results actually. So I am kind of questioning my actual personal experience". P10 found D1 to be a "definitely reassuring text" and P12 concurred, saying the "whole paragraph was definitely comforting,

especially since it's research-based encouragement".

F3: Users Are Unlikely to Click on a Citation Link but They Value the Added Credibility.

> "I think I would not click on links to research articles but maybe news articles, for example, Globe and Mail..." – P1

Participants expressed mixed reactions to being presented with a link to a research article (D4). Most of our interviewees were in their second year of undergraduate studies, and likely had limited exposure to reading research articles. While some students such as P4 observed that "it's an interesting study and it's really recent from 2018" and were able to summarize the key takeaways after quickly skimming through the abstract, several participants such as P1, 2 and 11 appeared confused by the complexity of the writing. P1 noted that they would prefer to read news articles: "I think I would not click on links to research articles but maybe news articles, for example, Globe and Mail... as long as they have a citation at the bottom that links to something credible". The last part of their comment indicates that even if they may not necessarily want to read a research article, they value having a citation link.

F4: Structuring Content as Listed Instructions May Help with Recall.

> "I will remember D2 because you have (1), (2), (3)...and with numbers, it's easy to remember. So this would be much more useful for me during the exam." – P5

D2 had some notable structural differences compared to D1, such as being separated into multiple paragraphs (despite being roughly the same overall length) and having concrete instructions listed as numbered points. We found that participants such as P5, 6 and 12, noticed these structural differences and commented on them without being explicitly prompted. P6 said "I do like how it's put into these strategic points" and shared that it's harder to remember things when you are stressed. P5 mentioned that having ideas as a numbered list may help them remember the points during exams.

F5: Stressed or Lonely Users May Find Comfort in a Short Talking-head Instructional Video.

> "...when you are feeling extremely stressed, you don't even have the mind to read a paragraph." – P8

We wanted to see whether the presentation modality (i.e. video vs text) made a meaningful difference with regard to how participants perceived the message. In D3, a faculty member read the instructions from D2. P8 observed that "when you're feeling extremely stressed, you don't even have the mind to read a paragraph", and that "you can always finish a video that's only 40 seconds". They also commented on how if the message is "from your professor or examiner, you will definitely watch a comforting message like this" leading to "a better effect compared to any text-formed message". Adding to P8's observation, P10 noted that the video modality felt better because "having someone explain it to you rather than just reading it makes a big difference as you feel like you're

being guided through the stress". In addition to being stressed, users may also prefer a talking-head video over text due to other factors such as feeling alone. For instance, P5 mentioned that "if I have been studying for several nights alone, and there's nobody talking to me, then I would prefer to watch a video like this. But if I'm reviewing for the exam with my friends and I don't feel lonely, I would prefer the text form". However, some students who are not particularly stressed such as P1 and P12, may not find a video appealing. P1 noted that having instructions in text form was easier to follow because a video had more "distracting" elements such as "his voice, his accent, and gestures", and P12 noted that reading allowed them to "go through it at my own pace",

F6: Writing May Make Non-native English Users More Conscious of Grammar and Word Choice.

> "When I'm typing, I'll pay more attention to getting the right word, having the right grammar, and finding a better word to explain." – P3

When discussing the two reflection prompts in D5 and D6, we found that several participants such as P2, P3, P5, and P7 preferred speaking out loud over typing because they felt more comfortable sharing rough ideas without worrying about grammar and word choice. P2, who was an international student whose native language is not English, noted that "when I type, it is kind of everywhere... I will also choose to correct my grammar and will have to look up certain words". In contrast, when describing speaking, they said "I will concentrate more on new ideas... I will just speak them... if I am writing them down, I may ignore the new ideas because I will (only) write down the things that I'm sure of". Similarly, P5 mentioned that "when I'm typing, I would think in my first language, Chinese, but I may have trouble translating my thoughts into written English. But if I'm speaking, I won't have time to translate things, so I would just choose some other easy way to say what I am thinking". Opting for less complex vocabulary may be tied to worries about being judged when writing, as hinted by P7's comment: "I feel like you wouldn't be criticized that much for talking incorrect [sic] compared to writing poorly". Elaborating on the latter half of P2's comment about only writing ideas that they are sure of, P6 noted that when talking out loud, they are more likely to share their "inner voice" which has "much more content" that gets lost when typing.

Students who felt more comfortable with writing, such as P8, P9 and P11, did not feel like they would share different ideas when speaking vs typing and described some advantages of the written form. For instance, P8 noted that they would focus on the main ideas, leaving word choice and grammar checks for the end. They also mentioned that "writing something down and actually considering your word choice would be a better way to remember everything". P11 preferred typing because they could "change" their words and "be more organized".

2.5 Study 2: Large-Scale Randomized Field Experiment

To further evaluate our design, we conducted a large-scale randomized field experiment where we looked at the impact of the intervention on exam performance, and whether there were any differences in terms of how it affected students of different genders and year of study. The results are summarized in figure 4.8.

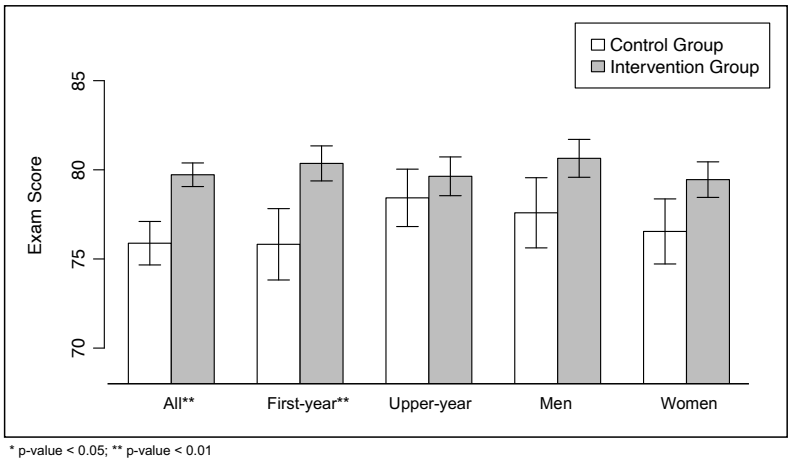


Figure 2.2: The differences in average exam scores between the control group and the intervention group at several levels. From left to right, we do the comparison within all data, first-year students, upper-year students, men, and women. The error bars show the standard error of each group.

2.5.1 Intervention Deployment

We embedded the stress intervention into an online course activity (as an optional section) that was distributed to students during the fourth week of class, about 10 days before an upcoming exam. This activity was one of five graded activities that we asked students to complete during the term, and students were given 2% of their final marks after their completion. The other activities were unrelated to stress reappraisal and contained multiple parts, meaning that our intervention had to compete for students' limited time and attention. The inclusion of those other activities may add more noise to the analysis of the stress intervention effect; however, we intentionally randomized those activities and parts separately to minimize the potential for inducing a fake effect as much as possible.

Participants

The intervention was deployed to 1283 undergraduate students enrolled in an Introduction to Programming course at a large research-intensive post-secondary institution in Canada in the Spring 2020 semester. 59.8% of the sample were first-year students, 22.8% second-year students, 9.3% third-year students, 5.3% fourth-year students, and 2.8% students in their fifth year or higher. They came from various disciplines including the physical sciences, natural sciences, social sciences, life sciences, humanities, and commerce. 54.6% of students took the course to fulfill program requirements, while 20.0% took the course as an elective. An additional 25.5% of students took the course to fulfill general education requirements or for other reasons. The gender identity of students in the class was 46.5% male and 51.4% female; 0.2% of students specified another gender, while the remaining declined to answer. Student demographic data was collected from a voluntary survey administered mid-semester. Note that all questions were voluntary, so a different number of students may have answered each demographic question.

Randomization

Students were randomly assigned to either a control group, which did not receive the intervention or a treatment group, which received the intervention. Those who were in the treatment group saw the core stress-reappraisal message and a random combination of one or more variants of additional content illustrated in Figure 5.1. The randomization was automatic and dual-anonymous. Neither the instructors nor the students had access to the assignment policy. We chose a 2:1 split between treatment and control due to ethical concerns and felt that it was not necessary for us to exclude half the students from the potential benefits of our intervention, given our large sample size. Balancing treatment and control is usually preferred because they result in a higher power. However, given that our sample size was sufficiently large, this was not a concern for us.

Data Cleaning

The intervention was delivered to 1284 participants in total. 1014 of them clicked the intervention link and 931 participants completed it. 92 out of those 931 participants either dropped the class or did not finish the midterm, leaving us with 839 participants. Of those 839, 175 finished the intervention late, i.e., after the midterm, and so we had to exclude them from our study because they did not follow our experimental protocol. We allowed late students to access the activity after the end date of our experiment due to ethical considerations. We wanted to give students the opportunity to receive a grade for completing the activity even if their data was not useful to us for our research purposes. This gave us our final count of 664 participants in the cleaned dataset.

Quantitative Analysis

To test whether the treatment group outperformed the control group, we used one-sided independent samples t-tests, with midterm scores as our dependent variable. We also constructed linear regression models to see if adding one or more sub-treatments affected the intervention outcome.

2.5.2 Field Experiment Results

Our quantitative analysis results are summarized in figure 4.8.

2.5.3 Effects on Exam Scores

We first compared the treatment and control groups by conducting a one-sided independent samples t-test. As shown in table 2.1 and figure 4.8 ('All, control group' vs. 'All, intervention group'), the stress-reappraisal intervention improved students' midterm scores significantly by 3.8 % (p-value = 0.003, Cohen's d=0.25).

2.5.4 Result Validation

To verify that we did not randomly assign more higher-performing students to the intervention group, thereby inducing a fake effect, we used a bootstrap approach to estimate the probability of observing a similar event by chance. More specifically, from the 664 students in our cleaned dataset, we randomly placed 203 participants into one group (so that the number matches what we actually have in the control group) and put the remaining participants in another group. Then we calculated

Table 2.1: Summary of sample sizes, group means, Cohen's d values and p-values from t-tests conducted between different groups. Note that there are 113 students who did not indicate their genders and 107 students who did not indicate their school year.

Group	Sample size	Group mean	Effect Size	Cohen's d	p-value
All, control group	203	75.88	3.84	0.252	0.003**
All, intervention group	461	79.72			
First-year students, control group	94	75.82	4.54	0.283	0.022*
First-year students, intervention group	212	80.36			
Upper-year students, control group	80	78.43	1.21	0.085	0.27
Upper-year students, intervention group	171	79.64			
Men, control group	77	77.59	3.05	0.208	0.09
Men, intervention group	155	80.64			
Women, control group	93	76.54	2.91	0.184	0.08
Women, intervention group	226	79.45			

* p-value <0.05; ** p-value <0.01

the difference between the average of these two groups and repeated this process 10,000 times to calculate the chance that the absolute value of such a difference is greater than what we observed in our true experiment, which was 3.84%. We found that the chance of this happening is less than 0.3%.

2.5.5 Subgroup Differences based on Gender Identity and Year of Study

Table 2.2: Results of the linear regression model where midterm score is the dependent variable and the predictor variables include: intervention, upper-year, and the interaction term of upper-year with intervention. The value of "Intervention" is 1 if the participant was in the treatment group, and 0 otherwise; the value of "upper-year" is 1 if the participant is an upper-year student, and 0 otherwise; the value of "upper-year and intervention" is 1 if the participant is a first-year student and was in the treatment group, and 0 otherwise.

	Estimate	Standard Error	p-value
(Intercept)	75.82	1.58	<0.0001
Intervention	4.54	1.89	0.017*
upper-year	2.61	2.32	0.26
upper-year and intervention	-3.33	2.80	0.24

* p-value <0.05

We also analyzed the differences in treatment effects based on gender identity and year of study. As shown in table 2.1 and figure 4.8, the stress-reappraisal intervention had a significant effect on first-year students (p-value = 0.02), while no significant effect was observed for upper-year students (p-value = 0.27). Moreover, the effect of the stress reappraisal intervention across different gender identities is quite similar: the average improvement among men was 3.05. Among women, it was 2.91. On the other hand, the effect of our intervention on first-year students was 4.54, which is considerably higher than that for upper-year students, where we observed an increase of only 1.21. These sub-group results, however, are mostly not statistically significant as we split the data into smaller subsets and more than 100 participants chose not to inform us of their gender identity or year of study.

We placed our data in a linear regression context to further examine the significance level of the intervention effect in first-year versus upper-year groups. The output variable for the linear regression model is the midterm score, and the input variables are 'received intervention', 'upper-year', and 'upper-year and received intervention'. The results are not significant, and hence we do not have enough evidence to claim that the effect of the intervention is significantly different in the two groups.

2.5.6 Relative Effects of different variants

To assess the relative effects of the six different design factors, we conducted a linear regression where the midterm score is the output variable and the input variables are the six design factors: D1, D2, D3, D4, D5, and D6. These factors indicate whether the participant received a specific type of intervention. The regression was conducted only on the data of the intervention group, and the estimates measure the *add-on effect* of the six design factors, and not the independent effect of each factor as we cannot separate them from our main intervention by design. For example, the interpretation of the estimate of D1, which is -0.03, is: given that a student entered the intervention group and received our main stress intervention, the average *additional* influence on their midterm score if they also received the design factor D1 is -0.03. As shown in table 2.3, we found that including the design factor D4 can significantly reduce the effects of stress reappraisal intervention (p-value = 0.015), while all other design factors show no significant individual effects within the treatment group.

Table 2.3: Results from the linear regression model where the midterm score is the dependent variable and D1 to D6 are predictor variables. The value of 'Di', where i = 1, 2, 3, 4, 5, and 6, is 1 if the participant is in the intervention group and received the design factor i as one of the interventions. We only look at data in the intervention group and all the estimates are measuring the add-on effect of the six design factors upon the main effect. Each participant can receive multiple design factors.

	Estimate	Sample Size	Standard Error	p-value
(Intercept)	81.05	664	1.79	<0.0001
D1: Summary Text	-0.03	233	1.34	0.98
D2: Listed Instructions	-0.47	229	1.33	0.72
D3: Video Instructions	-0.88	226	1.33	0.51
D4: Link to Paper	-3.27	230	1.33	0.015*
D5: Self-Explanation	1.68	241	1.34	0.21
D6: Note to Self	0.21	231	1.32	0.87

* p-value <0.05

To check whether the number of interventions in a certain treatment influences the treatment effect, we analyze the *add-on effect* of including more design factors to the effect of the main intervention. The linear regression result is shown in table 2.4, where the p-value is 0.334. We also conducted an ANOVA test of the midterm scores among participants in the intervention group that received D0 through D6. In this ANOVA test, we set the number of design factors received as a categorical variable, so as to see if a certain number of sub-interventions have a statistically significant difference among other possibilities. The p-value of the test was 0.43, which is not significant. To ensure our analysis was robust, we repeated the regression and the ANOVA analysis on data that does not count D4 as a design factor or only on students who were in the intervention group but

did not receive D4. We removed D4 from our analysis because it had a statistically significant and negative effect. However, in any version of that analysis, such as with D4 included, the result is not significant.

To conclude, we found our main intervention had a positive and statistically significant effect on students' test performance, and that we should be cautious about including potentially counterproductive design factors such as the reference link in the intervention.

Table 2.4: Results from the linear regression model where the midterm score is the dependent variable and the number of interventions is the predictor variable. We only use data from the intervention group. The number of interventions is the total number of design factors received by the participant.

	Estimate	Standard Error	p-value
(Intercept)	81.32	1.78	<0.0001
Number of interventions	-0.53	0.55	0.334

2.6 Discussion

In the following section, we begin by summarizing the key findings of the two studies and underscore their relevance to intervention designers. Then, we elucidate design implications tied to elaboration, modality, prompts, timing, and target audience. Finally, we describe opportunities for future research and outline the limitations of our work.

2.6.1 Key Findings

Online *eustress interventions* that help students embrace the positive aspects of stress to focus and perform better on exams have great potential, but it is less clear how designers and researchers can make specific decisions about the digital delivery of such information to students. To address this issue, we explored six design factors (D1-D6) that embody components for online interventions that are *brief*, *voluntary*, and *scalable*, and work by reinforcing a core reappraisal message (D0) through a layered approached. We systematically evaluated our designs through two studies:

In Study 1 (described in section 2.4), we thematically analyzed in-depth interviews with 20 participants to derive six findings (F1-F6) that underscore the need for reinforcing the core message using various sources and methods. We found that a short stress-reappraisal message, while powerful, isn't sufficient for convincing users (F1) and that users valued *research-based* guidance and encouragement (F2). Users also shared that while they were unlikely to click on a paper citation link, they often considered the *mere presence* of the citation as an indication of increased credibility (F3) for the stress reappraisal idea. We also saw how content structured as listed instructions was perceived favourably by participants because they thought listed information would be easier to recall during the exam (F4). We observed noteworthy subjective differences in users' perspectives on the same information being presented as text or video. Participants expressed a preference for instructional videos in situations where they were stressed or lonely because they valued the comfort and guidance afforded by a talking-head video from their instructor (F5). Finally, we identified some differences in how users express their ideas when writing or talking out loud, such as focusing more on grammar and word choice when typing, and expressing more ideas when speaking (F6).

In Study 2 (described in section 2.5), we found that our design had a significant positive effect on exam scores ($p = 0.003$, $d = 0.252$) in a large programming class. Our subgroup analysis indicated a significant effect for first-year students but not for upper-year. We did not detect significant gender differences. In F3 from Study 1, participants expressed that they found the presence of explicit paper citations to be more credible. In contrast, in the analysis of the relative effects of our design factors in section 2.5.6, we found that including the paper citation can significantly *reduce* the effects of stress reappraisal intervention (p-value = 0.015), while all other design factors show no significant individual effects within the treatment group. Therefore, we caution instructional designers when considering whether to include similar design components which involve external links. We suspect that participants who clicked the link may have become distracted. Furthermore, the perception of increased credibility does not automatically translate into increased internationalization of the reappraisal message. The significant and positive impact on exam performance was especially surprising and remarkable because participants spent on average, *only three minutes* on the intervention. We validated our results using a bootstrap approach described in section 2.5.4.

2.6.2 Design Implications

In this section, we synthesize our findings into five design implications that are tied to the value of additional elaboration, presentation modality, reflection prompts, timing, and target audience.

Additional Elaboration: Providing More Information versus Preserving Focused Attention

Firstly, a key question that designers should consider is the trade-off between adding more information to ensure students understand and apply the concept (explanatory context in D1, concrete suggestions in D2), and the limits on attention and the potential visual clutter, given that students spent an average of only 3 minutes on this intervention. Our findings from the first study on exploring user perspectives showed the value of having additional elaboration as many students found it useful to have the research-based explanatory elaboration from D1, and some who were more skeptical appreciated seeing the citation from D4, even if they may not necessarily click to read a long research paper. Students also felt that having actionable instructions as a list in D2 would make it easier for them to remember how to apply the intervention during the exam. Therefore, instructional designers wishing to incorporate similar mindset interventions that are counter-intuitive can take a multi-pronged and layered approach to convince their students of the idea by making ample use of research-based encouragement in their content.

Presentation Modality: Delivering the Same Text as a Simple Video

Secondly, a simple video reading the same exact text from D3 is valued by students because they can *see* a person explaining why they should believe in the message and how to apply it. Designers might consider including such simple conversational videos as students shared that they felt guided through the stress reappraisal information when presented in video form, especially when feeling stressed (F5). This latter point about feeling stressed is important because, in situations where participants may not be predisposed to consuming content in a certain medium, intervention designers can help

learners by offering the same information in a different form. This can also have the added benefit of improved accessibility.

Reflection Prompts: Internalizing and Applying Eustress to Everyday Life

Thirdly, one insight was that the reflection prompts guide students to think through how to apply the eustress intervention to everyday life, and that is certainly possible. A reflective learning approach helps in addressing a key challenge for instructional designers, which is to help learners retain and apply intervention content after finishing the activity. Therefore, we encourage intervention designers to consider how to augment their activities with reflective exercises, especially near the end of their interventions.

Intervention Timing: Delivering Interventions when Learners are Most Receptive

Moreover, during the interviews, certain students mentioned that they were less likely to go through such activities closer to the exam because, around that time, they would rather focus on exam content as they are worried about finishing the syllabus. Such comments signify the need for intervention designers to consider the tradeoff between delivering the information closer to the exam so they remember the intervention content, and the need for students to focus on other things that compete for their limited time and attention.

Target Audience: Eustress interventions may be more helpful to first-year students

Finally, in the analysis of our field experiment results in Study 2, the intervention had an overall positive significant effect on exam performance. However, when we conducted the subgroup analysis for first-year vs upper-year students, we only observed a significant effect for the first-year group. This finding is in alignment with prior work on stress reappraisal [32]. Certain comments from participants hinted at some plausible reasons as to why we observed this difference: students in their first year may be more stressed as they just transitioned from high school to university education, which can be challenging for many. Furthermore, first-year students may also be more receptive to adopting new exam strategies whereas students in more senior years may have already formed strong opinions on how to best tackle exam stress. Intervention designers can therefore benefit from analyzing subgroup differences when evaluating their interventions, such as any difference between first-year and upper-year students.

2.6.3 Future Work

The content of our intervention as fully described in section 2.3.2, did not contain any references to course-specific materials, and as such can be easily generalized to other courses, and thus, this work can inform several directions for future research in the expansive design space for online eustress interventions. Firstly, from our exploration of the value of additional elaboration, we saw that users like articles or activities that are framed as being "evidence-backed" but may be less likely to explore the papers linked in the citations. We also found in the field experiment that the citation link had a significant and negative add-on impact in our analysis of the relative effects of each design factor. Future work can explore the specific contexts in which such an observation holds or does not hold, and whether certain student populations are more or less inclined to accept a research-based

framing. Secondly, when varying presentation modalities, we saw the promise of offering the same content in a different format. Our video consisted of one individual delivering a particular message, and it is hard to know what properties would generalize to other settings. Future work can explore what dynamics affect how students react to reappraisal messaging in the video form, such as how students respond to pop-up messages on eustress when presented as in-video prompts [313] for them to do something during or after they watch the video, as well as other kinds of visualizations of the reappraisal message, such as the presence of other agents in the video including past students instead of an instructor. Thirdly, we saw how adding reflection prompts at the end of mindset interventions may help learners retain the message. Future work could examine if asking students to explain by recording a voice message helps them internalize the reappraisal information. Students could also be asked to record a voice or a video message to *themselves* that could then be sent by email or text at a particular time. These examples illustrate how our work can help suggest future directions for designs that explore how we might send these messages at the right moment, in the right format in a way that engages people in receiving reappraisal messaging to change their mindset and behaviour through technology-mediated online interventions. Finally, a more detailed exploration of how such online interventions can be integrated within different contexts such as in-person vs remote learning can further inform the design of future eustress interventions.

2.6.4 Limitations

Replications of our approach can be further strengthened by including pretest and post-test performance measures, i.e., having student test scores from an initial test compared to performance on a second test with the intervention in between. Furthermore, if there are multiple opportunities for testing performed at different time points after the intervention is deployed, those tests can help inform whether there is a sustained impact of such interventions or if it is a single impact that needs boosting. In our case, the final exam was cancelled due to COVID-19 when we deployed our intervention. As a result, we only had the midterm scores as a post-test measure.

It is possible, though not likely, that we randomly assigned higher-performing students to the intervention group. However, to check for this, we used a bootstrap approach and found that the chance of getting a similar or more significant event by randomly splitting students into two groups is less than 0.3%. It is also possible that higher-performing students have less stress and therefore perform better on exams, and stress may not be equally distributed across all performing levels of students.

Our intervention was accessible through a web page in a field setting. This setup maximized the ecological validity of our results. At the same time, it led to complex implementation issues that we tried to minimize, such as accounting for page refreshes by setting up our randomizer such that if participants' reloaded the activity, they would remain assigned to the same experimental group (i.e., either control or intervention) through browser caching. However, a small percentage of participants (1.7%) loaded and completed the activity from more than one device (e.g., by doing it from their phone after seeing the announcement and then doing it again on their laptop). These participants were inadvertently labelled as being in both the control group and the intervention group by the randomizer. However, because they saw the intervention, they were of course no longer part of the control. We factored this into our analysis by treating those participants as being in the intervention group. We reassessed whether our decision to leave this small percentage of

participants in the intervention group induced an artificial effect by rerunning the analyses with those participants excluded. The results remained similar to our current analysis, e.g., the main effect remained significant ($p = 0.018$).

2.7 Conclusion

Our work presents an approach for designing effective interventions for shifting student attitudes of exam *stress* towards *eustress*. We explored the design space for online stress reappraisal interventions that were *brief*, *voluntary*, and *scalable*. We conducted two studies to explore design dimensions that varied in the levels of elaboration, presentation modality, and the use of reflection prompts. We instantiated these dimensions using six design factors (D1-6) that reinforce a core reappraisal message (D0) and evaluated these factors through both semi-structured interviews with 20 students, and a randomized field experiment deployed to over 1200 students in a real-world programming course. The interviews yielded practical insights into the relative importance and impact of factors on different students in different contexts, such as when text versus video presentation might be effective, what kinds of additional information are compelling versus burdensome, and what kind of interface prompts for students are more or less impactful in helping them retain and utilize the reappraisal messaging. These dimensions can be used to generate an expansive space of design variables that can be explored in future work: from the modality of videos that incorporate various visualizations, reflective activities such as students' recording voice and video messages for their future selves, to adaptive delivery systems that target message timing to the right moment and mental state. Although students only spent an average of 3 minutes on the intervention, our field experiment revealed that combinations of the six design factors were powerful enough to boost the class average exam score from 76% to 80%, which is equivalent to an increase from a **B** to an **A-** in the grading scale used by the course where we deployed our study.[1]

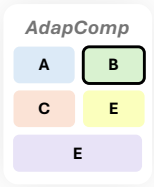
2.8 Acknowledgements

We are grateful to the students who participated in this work. We would also like to thank Andrew Peterson at the University of Toronto, Gerry Chan at Dalhousie University, and René Kizilcec at Cornell University for their input on the early drafts of this manuscript. This work was partially supported by grants from the Office of Naval Research (N00014-18-1-2755, N00014-21-1-2576), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-06968), and the National Science Foundation (2209819).

[1]Note: grading systems vary by province and institution in Canada, and may be different from US schools. The grading scale relevant to this study is used by certain institutions of higher education in Ontario.

Chapter 3

The AdapComp Framework



> The HCD principle is to avoid specifying the problem as long as possible, instead iterating upon repeated approximations.
>
> *Don Norman [242]*

Research Context: In the previous chapter, we explored how investigating multiple alternative designs in parallel can be used to develop powerful resources that improve learning outcomes. In addition to *parallel* exploration, another essential element of experiment-inspired design is *continuous* experimentation. This chapter focuses on that goal, addressing the second of our four primary research questions:

> *Q2: How might we instrument educational platforms to accelerate the use of research to improve students' experiences through continuous experimentation?*

Our answer to this question is the AdapComp/MOOClet[1]framework which enables instructors and researchers to set up digital learning platforms for continuous experimentation by treating digital content as collections of modular components that can draw from a growing pool of design variations. Instructors can connect their platforms to the framework even before formally specifying experiments, and can add to the version pool as new ideas arise. We used this framework in various interventions within and outside of U of T, and used it in a set of deployments to win the million-dollar XPRIZE digital learning challenge in 2023.

Related Publication:

Mohi Reza, Juho Kim, Ananya Bhattacharjee, Anna N. Rafferty, and Joseph Jay Williams. 2021. *The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses.* In Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21). Association for Computing Machinery, New York, NY, USA, 15–26. https://doi.org/10.1145/3430895.3460128

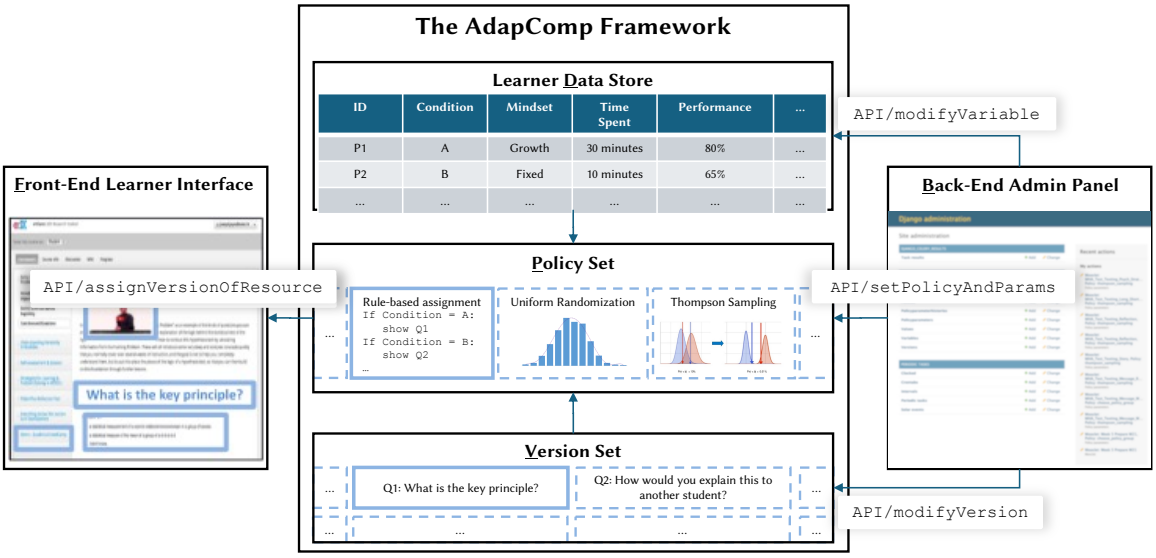


Figure 3.1: The AdapComp/MOOClet framework architecture consists of the Learner Data Store, Policy Set and Version Set. These components serve as an abstraction layer between the front-end Learner Interface and the Back-End Admin Panel for instructors, researchers and developers, who can interact with the framework via API calls.

Abstract: How can educational platforms be instrumented to accelerate the use of research to improve students' experiences? We show how modular components of any educational interface – e.g. explanations, homework problems, even emails – can be implemented using the novel AdapComp/MOOClet software architecture. Researchers and instructors can use these augmented MOOClet components for: (1) **Iterative Cycles of Randomized Experiments** that test alternative versions of course content; (2) **Data-Driven Improvement** using adaptive experiments that rapidly use data to give better versions of content to future students, on the order of days rather than months. A MOOClet supports both manual and automated improvement using reinforcement learning; (3) **Personalization** by delivering alternative versions as a function of data about a student's characteristics or subgroup, using both expert-authored rules and data mining algorithms. We provide an open-source web service for implementing MOOClets that has been used with thousands of students. The AdapComp/MOOClet framework provides an ecosystem that transforms online course components into collaborative micro-laboratories, where instructors, experimental researchers, and data mining/machine learning researchers can engage in perpetual cycles of experimentation, improvement, and personalization.

3.1 Introduction

Even after online courses are deployed to students, many instructors wonder about how to identify better versions of course content. For example, if instructors and researchers could add alternative explanations of key concepts, they could experimentally compare which version is more helpful to

[1]We initially called this framework MOOClet, but in more recent iterations, as we applied it to contexts beyond education, we adopted a more general term AdapComp, short for Adaptive Component.

students [367, 368]. Data from initial experiments could be used to enhance the experience of future students [368], as well as inspire ideas for new explanations to experiment with, producing iterative cycles of data-driven improvement. In addition, data about how alternative explanations benefited students with different characteristics could be used for personalization, such as giving an explanation of Type A to students with lower prior knowledge versus Type B to those with higher prior knowledge [367, 339]. How can educational platforms better facilitate such iterative experimentation, data-driven improvement, and personalization?

Our answer is the AdapComp/MOOClet software architecture, which augments front-end components of educational interfaces with back-end APIs and databases. These are designed so course content can be flexibly adapted, through the process of instructors and researchers exploring new ideas, data, analysis and algorithms. We refer to any interface component augmented with this back-end architecture as a MOOClet. Figure 3.2 and the following usage scenario illustrate what a MOOClet is and the affordances it provides: An instructor has a MOOClet implemented which delivers an explanation of standard deviation on a particular lesson page in edX by instantiating a MOOClet back-end and link it to the front-end webpage. The motivation for the instructor to use the MOOClet is that they anticipate that future research could be conducted to improve the explanation (or other webpage content). Although there is not yet a formulated plan for what ideas to test and the MOOClet simply delivers the single original explanation, it leaves the door open for future improvement.

Imagine that students' questions and comments on the discussion forum later suggest an alternative explanation for standard deviation. This is added to the MOOClet, and the probability of assigning the two explanations is manually set to [90% Original,10% New] at first out of caution but with no red flags it is changed to [50% Original, 50% New]. The instructor collaborates with a researcher to evaluate the two explanations using quantitative metrics (e.g., ratings of explanations, accuracy in solving related quiz questions) as well as qualitative feedback (students' comments and questions on the explanations). The instructor comes up with a third explanation based on this data, and after another experiment with probabilities set to [33%, 33%, 33%], decides it is better and adopts it for the foreseeable future by setting probabilities to [0%, 0%, 100%]. The MOOClet enables this cycle of iterative experimentation and data-driven improvement to be repeated anytime in the coming years as new ideas arise or data emerges about how to better explain.

The MOOClet also enables automation of data-driven improvement. Logging code can be used to send the MOOClet metrics like student ratings of explanations. The next time new explanations are proposed, to avoid manual analysis, the course team could use a built-in widely-used multi-armed bandit algorithm for adaptive experimentation [353, 210] to automatically analyze explanation ratings and change the experience for future students. The algorithm automatically modifies the probability of assigning an explanation, using Bayesian analysis of the probability the explanation is the highest-rated. As explained in Use Case 2, course teams can access a user-friendly dashboard built on the MOOClet APIs, to monitor the algorithm's behavior in case they want to intervene [368].

The MOOClet also enables personalization. As a more diverse range of students take the course, survey data suggests that changing the vocabulary and examples of an explanation could make it more helpful to students from different countries. The MOOClet enables the course team to work with researchers to either: (1) run an experiment to test the impact of alternative phrasings on

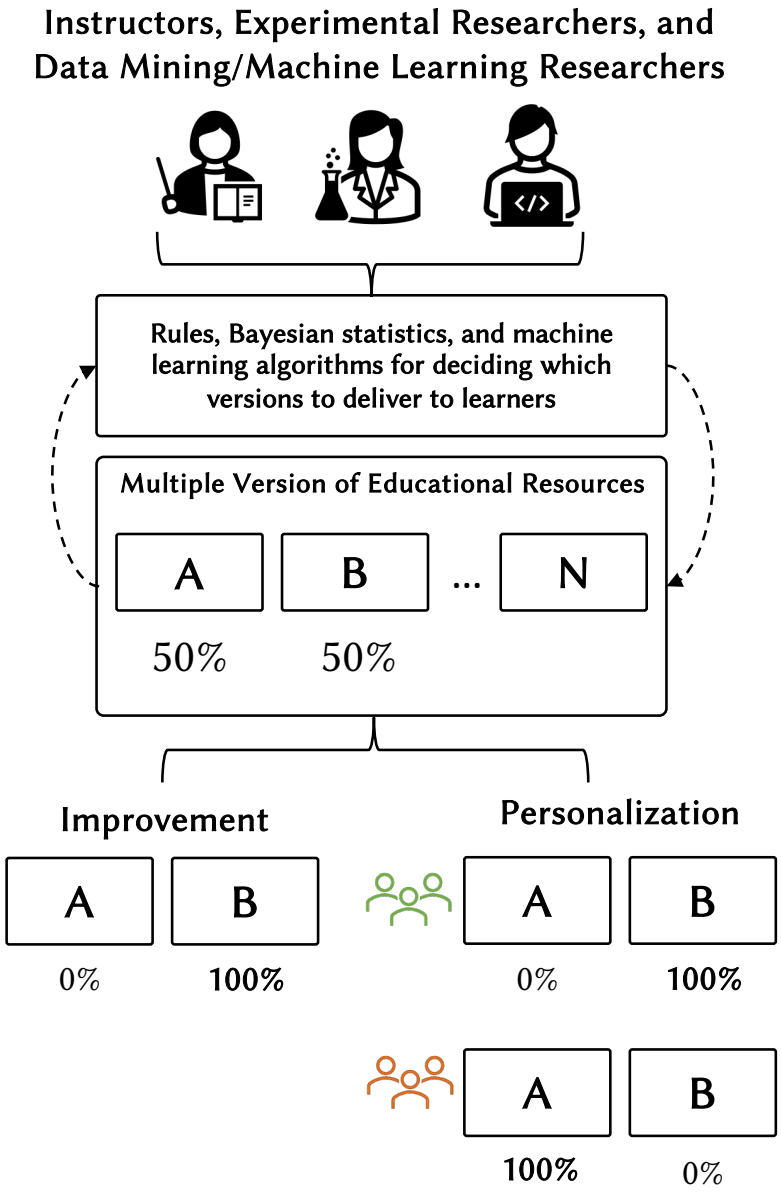


Figure 3.2: The AdapComp/MOOClet Framework enables Instructors, Experimental Researchers, and Data Mining/Machine Learning Researchers to collaboratively conduct A/B comparisons to improve and personalize educational resources in online courses.

different subgroups of students; (2) write `IF-THEN` rules that give certain explanations to students from certain countries; (3) apply a range of algorithms for personalization and recommendation to optimize delivery of different content based on a more complex learner profile.

This usage scenario illustrates how implementing a course component as a MOOClet allows for perpetually iterative cycles of experimentation, data-driven improvement, and personalization, for a range of future research where the exact ideas, data, analysis or algorithms may not be specified in advance.

There are many challenges in achieving the usage scenario using current platforms, including: (i) the use of one-off software implementations to answer a *particular* pedagogical question in a *particular* Learning Management System (LMS), (ii) difficulties associated with making changes to existing resources once a course goes live, (iii) limited flexibility in testing out ideas or using data in ways that were not part of the original study design, (iv) the challenge of combining and managing data from various sources.

The *technical contribution* of the MOOClet framework is a set of design requirements and a proof-of-concept web-service implementation that exemplifies how to architect software to enable the kind of research we envision. We describe how to implement different front-end components as MOOClets, provide a set of design requirements for the framework, and show how to improve and personalize resources over time using a three-fold abstraction layer consisting of a *Version Set*, *Policy Set* and *Learner Data Store*, as illustrated in Figure 3.1. Our implementation supports (i) easy switching between experimentation, improvement, and personalization of educational resources, (ii) the creation of flexible policies for deciding what to present, (iii) the use of machine learning algorithms and reinforcement learning to automatically analyze data and present higher rated resources to future students, (iv) the addition of new versions of resources at any point in time without major redeployment of front-end code, and (v) personalization of resources using data mining algorithms, as well as manual rule-based specification of which problems to give to learners with different characteristics.

To summarize, this paper makes the following contributions:

- The conceptual contribution that a unified software infrastructure can be used to architect modular components of educational interfaces to enable practical research using iterative experimentation, data-driven improvement, and personalization.
- Eight design requirements for the MOOClet Framework, a detailed description of its architecture, and an open-source web-service implementation that provides back-end databases and APIs that plug into multiple platforms such as MOOCs and Learning Management Systems.
- Real-world use cases of MOOClets demonstrating how to use them and when their use can bring together instructional improvement, experimental research, and data mining/machine learning algorithms.

3.2 Related Work

We consider related work on the implementation of online field experiments, accelerating improvement of educational resources using data from experiments, data-driven personalization, and common educational data standards and infrastructures.

3.2.1 Implementation of Online Field Experiments

Over the past few years, experiments and A/B testing have become ubiquitous in the technology industry, particularly in the context of developing user interfaces and marketing [86, 179]. Tools for experiments have been deployed in industry settings like website testing, where companies like Facebook and Microsoft have substantial resources to hire programmers to implement experiments and apply machine learning algorithms for product improvement and personalized recommendations. Tools like Planout [15] and Optimizely [323] have made end-user experimentation on websites possible by allowing users to define the logic of experiments, such as factorial designs and stratified sampling. However, there is far less functionality for randomized experimentation in most Learning Management Systems and MOOC Platforms, with some notable exceptions being platforms like edX [35] and ASSISTments [129]. A key advantage of the MOOClet framework is that it augments existing platforms with capacities for experimentation, and further provides functionality for data-driven enhancement and personalization.

3.2.2 Accelerating Improvement of Educational Resources using Data from Experiments

In industry settings, rapid use of data from experiments to improve products is a top priority [351], as A/B comparisons are seen more as a tool for product improvement than scientific research. In addition, machine learning algorithms like Thompson Sampling for the multi-armed bandit problem have been applied by tech companies to rapidly analyze data from randomized experiments, to try to accelerate the creation of better performing resources and experiences for future users [300]. Multi-armed bandit algorithms have also been used in some cases to improve educational experiences for students [60, 210]. Although such adaptive experiments have been applied with great success by leading technology companies, they introduce a range of complex issues concerning statistical analysis, which are active areas of research [274]. The aim of this paper and the MOOClet framework is not to solve all the challenges that arise from using adaptive experiments, but to make it easier for instructors and researchers to have the *option* to even conduct adaptive experiments for rapid data-driven improvement. This can also open up new opportunities for research on how to improve methods for adaptive experimentation in real-world contexts.

3.2.3 Data-Driven Personalization

There is a vast literature on using data for personalization in education, such as in Intelligent Tutoring Systems (ITS) [264, 63], as well as work on educational recommender systems [260]. Personalization can be understood as delivering one of several alternative experiences, as a function of data about an individual student – such as recommending one of several problems as a function of data about how a student performed on similar previous problems. Existing approaches to personalized learning in online courses tend to explore the use of specific factors such as competency [257], prior knowledge [171], or level of motivation [160]. The MOOClet framework, in contrast, is a *general* approach to personalization where *any* student characteristic or interaction with the learner interface can be used for personalization, as long it can be logged, and *any* algorithm, manual or automated, can be used to decide which versions to deliver to the learner interface. Furthermore,

the famework's approach to personalization is also *flexible* because at any point in time, these characterstics, interactions, and algorithms can be easily modified. Therefore, MOOClets can make it easier to apply existing approaches to new settings and then refine them over time.

3.2.4 Relationship to Common Specifications, Standards & Infrastructures in Education

How does the MOOClet framework relate to existing useful education standards surrounding data use? Efforts like MOOCdb [260] and xAPI [374] aim to provide a common format for data across multiple platforms or components of a platform, so that it can be readily understood by researchers, or used by specific tools. The MOOClet framework is not a *standard* for data, but an *architecture* for implementing educational resources that learners interact with, so as to enable a range of experimentation and improvement. So the Learner Data Store that a MOOClet accesses can be implemented using specifications from MOOCdb or xAPI, and the key value addition of the MOOClet is that such data could be easily used by multiple different algorithms or methods for experimentation, dynamic improvement, and personalization.

3.3 Design Requirements

The overarching design goal of the MOOClet framework is to implement the underlying software architecture for components of real-world learner interfaces so that they are "future-proofed" for a broad range of not-yet-specified research involving manual and automated methods for Experimentation, Data-Driven Improvement, and Personalization.

To operationalize this overarching goal and delineate a set of constraints and considerations for the framework, we define eight design requirements (D1-8). These are motivated by field observations, conversations, and interviews during the authors' experiences with conducting over fifty randomized experiments across a range of software platforms (Khan Academy, ASSISTments, edX, NovoEd, Moodle, Canvas), our review of existing software and tools for experimentation, and our characterization of the relationship between experimentation, data-driven improvement, and personalization.

D1: Enable *iterative experimentation* based on randomized A/B comparisons where successive versions of resources can be added or removed. To help users investigate the advantages and disadvantages of multiple resource versions, our framework must enable iterative A/B testing, where data is readily available for analysis, and where potentially better versions of resources can be easily added for investigation even after real-world deployment.

D2: Enable resource *improvement* using data on past learners. To help framework users choose between alternative resource versions, and converge towards the best version, the framework must collect data about past learners who have received alternative versions of the resource.

D3: Enable resource *personalization* using data about specific learners. To help framework users account for the heterogeneity in learning profiles and preferences between different students, as is typical in large and diverse online student populations, the framework must help users take into account different student characteristics to personalize which versions are delivered to them.

D4: Work with existing learner interfaces. To maximize user adoption, our framework must be compatible with existing learner interfaces and require only minimal, modular changes to the front and back end infrastructures of these interfaces.

D5: Support the addition, modification, and removal of resource versions at any point in time. Whenever new ideas arise, framework users must be able to add them into the system, and test them against existing versions. This should be possible at any point in time, as opposed to only before deploying a course, or after finishing it. For example, let's say in week 1, learners see version A of an explanation. In week 2, if users want to test a new version, the framework must allow them to easily deliver it to some subset of learners, and compare it with earlier versions.

D6: Support multiple methods for deciding how versions are delivered to learners. These include but are not limited to uniform randomization, weighted randomization, and personalization based on learner characteristics and their interactions with the learner interface.

D7: Support the addition, modification, and removal of policies for delivering resources at any point in time. Framework users should be able to change the methods being used to assign versions, and alter parameters such as the weights or probabilities of assignment and any rules used to select between versions.

D8: Support the continual addition of data from multiple sources for use in improvement & personalization. This ensures increasing access to unanticipated or not yet available sources of data, in order to inform both research and practical improvement.

3.4 MOOClet Architecture & Web Service

The design requirements motivate the MOOClet architecture. This section explains what the components of the MOOClet architecture are, how components interact with each other, and the APIs that must be available, as illustrated in Figure 3.1.

To preview, implementing a Resource in a front-end Learner Interface as a MOOClet requires that the Version assigned to a resource is obtained by an API call to the MOOClet back-end, which consists of a Version set, Policy, and Learner Data Store associated with a particular MOOClet. Specifically, the API call uses the MOOClet's associated Policy (rule/algorithm) to choose a Version from its Version Set, with the Policy having access to the variables in the Learner Data Store to choose Versions. Critical to the architecture is that there must be APIs for accessing, modifying, and adding to the contents of the Version Set, Policy Set, and Learner Data Store. We elaborate more on each component below.

Open-Source Web Service Implementation. We instantiated the architecture in a web service for using MOOClets that deliver text and HTML Resources. We used the Django (python-based) framework for web applications, to provide: (1) Classes that allow the creation and modification of SQL database objects to instantiate particular MOOClets and associated entities (e.g. Version Sets, Learner Data Stores); (2) Appropriate RESTful APIs (see example calls in Figure 3.3); (3) a graphical user interface Admin Panel as an alternative to the APIs. The *MOOClet Use Cases* section used this MOOClet web service, and www.mooclet.org provides details on interacting with our web service and/or implementing one's own. (4) A number of policies, elaborated on in the *Examples of MOOClet Policies* section.

We now elaborate on the key concepts:

Learner Interface: The front-end educational-interface that displays the content a learner will

interact with (e.g. Canvas, edX, Khan Academy, Coursera, Qualtrics, Tools using Learning Tools Interoperability, emails, text messages, mobile apps), into which MOOClets are embedded.

Resource: A component of a Learner Interface that presents a Version of content, experience, or interaction to a learner (e.g. paragraph on a website, an explanation to a problem, an email sent to students, a video lesson, HTML code for a problem, a reflective prompt etc.) and is implemented as the front-end part of a MOOClet. The Version presented to a learner must be chosen via an API call to the back-end, to allow flexibility in adding Versions and changing Policy, rather than the typical approach of hard-coding these into code enmeshed to the Learner Interface, which is far more complicated to modify.

Version & Version Set: A back-end data structure that contains alternative Versions that are delivered via Resource in the Learner Interface. An API must be available for retrieving, adding, and modifying Versions.

Learner Data Store: A data structure that can contain a wide range of data and variables about learners that is useful for experimentation, improvement, and personalization. Variables tied to a learner are linked by an anonymous learner ID. These variables can include which Version a particular learner received, and an accompanying metric/dependent variable for evaluating the impact of the Version. Variables relevant for personalization include student characteristics, such as whether a learner got a previous problem right. Variables can also be added that represent new variables produced in interim statistical analyses, as well as algorithm parameters. An API must be available for retrieving, adding, and modifying variables.

Policy: A function for determining which version of a Resource is presented to a particular learner, from decision rules as simple as "assign with probability XX" to as complex as any range of algorithms. A Policy can have parameters, which can be modified by API. A Policy has API access to use variables from the Learner Data Store as input. An API must be available for changing the Policy associated with a particular MOOClet (or, relatedly, modifying a Policy's parameters) at any point in time.

Policy Set: A collection of all the potential Policies a MOOClet can use, which can be added to and extended.

Admin Panel: An optional interface for making it easier for users to interact with the components associated with a MOOClet, such as a graphical user-interface that provides access to the API calls previously specified.

MOOClet: We use the term MOOClet to refer to a particular implemented constellation of a Resource as a front-end component of an educational interface (e.g. text on a webpage), an associated Policy, Version Set, and a Learner Data Store with the appropriate APIs. A front-end Resource component is 'counted' as a MOOClet if and only if it is linked to a Version Set, Policy, and Learner Data Store, and if these have been implemented using the specified architecture and APIs. All of the following must be true: the content displayed in the MOOClet interface component is selected by an API call which uses an associated Policy (from the Policy Set) to chose a Version from the Version Set; the Policy has API access to data from the Learner Data Store in selecting versions; API calls can be used to modify and access data in the MOOClet's associated Version Set, Policy, and the Learner Data Store.

The reason we introduce the novel term *MOOClet*[2] is to be precise in labeling educational

[2]The stem MOOC and diminutive "let" were chosen because a natural use of this approach is to design the

resources implemented using this architecture, as discussions with potential users revealed they often believe an interface component enables what a MOOClet does, but if the architectural constraints are not met there is inevitably some capacity missing. For example, a component can be randomized, but code reimplementation (rather than API calls) is required to enable personalization. Or a component can be adapted by a single algorithm for adaptive experimentation, but switching to using an alternative algorithm (and sometimes even changing an algorithm parameter!) requires redeploying the Learner Interface code.

MOOClet Users: Instructors, Experimental Researchers, and Data Mining/Machine Learning Researchers can use MOOClets to conduct experiments, analyze data, or use algorithms for data-driven improvement and personalization. The people who typically have programming skills needed to implement experiments, data pipelines, and algorithms, will also connect the MOOClet architecture to existing educational platforms, and initialize MOOClets for specific applications. The architects and developers for educational platforms (e.g. Canvas, edX) and local university IT support staff can also use and adopt the MOOClet architecture.

3.4.1 Examples of MOOClet Policies

We highlight some of the Policies currently implemented in the Web Service, while noting that others can be added to the policies file in the github repository (tiny.cc/githubmooclets).

A ***(Weighted) Randomization*** Policy that takes as parameters a list of probabilities of assigning any learner to each version. These probabilities can be uniform or weighted. For example, if there are two versions, and the weights are set to [50%, 50%], each version will have a 50% chance of being delivered. This is the uniform variant. An example of the weighted variant would be [20%, 80%]. In this case, version A will have a 20% chance and version b will have 80% chance of being delivered to the learner. This policy can also be used to give one version that is deemed to be the best to everyone, with probabilities like [0%, 100%]. These parameters can be updated at any point via API, so that manual updating allows for improvement by giving resources that seem effective to more future students.

A ***DynamicRandomization*** Policy, which takes as a parameter an outcome variable from the Learner Data Store that it should choose versions in order to maximize. It does this using algorithms for solving multi-armed bandits from reinforcement learning, specifically a Bayesian algorithm called Thompson Sampling [49]. The use of Thompson Sampling for DynamicRandomization aims to optimize for an outcome variable that is available in the Learner Data Store (such as accuracy on subsequent problems), by trading off assignment to the different Versions of a MOOClet (exploring) against always assigning learners to the Version that produces the highest outcome (exploiting knowledge). Conceptually, DynamicRandomization can be understood as automatically reweighting randomization, where the probability of assigning a Version is the probability that it is the best Version (on the target outcome variable) based on the data collected so far. So every time more data becomes available, DynamicRandomization has an updated set of probabilities.

A ***(Weighted) Personalization*** Policy takes as parameters `IF-THEN` rules that specify how assignment of Versions to a learner depends on data in the Learner Data Store, such as a learner's

software underlying components of MOOClets. However, it will become apparent that the approach is not restricted to MOOCs, and the requirements specification can be used to ensure dynamic experimentation and personalization in any digital educational resource. In fact, it can be used for a range of user-facing software, from websites to emails to mobile apps.

characteristics. The THEN clause can also provide a set of weights/probabilities of assigning different versions.

An ***External*** Policy which assigns Versions by using an External Eolicy via API, sending out relevant variables from the Learner Data Store. Use case 4 uses an external Policy to do problem recommendation based on applying Bayesian Knowledge Tracing to learners past behaviors. A wide range of multi-armed (contextual) bandit and other reinforcement learning algorithms could be used as External Policies.

Adding and Refining Policies: The web-service implementation GitHub repository is accessible by all MOOClet users. Any new Policy added by a data mining or machine learning researcher becomes available for instructors and experimenters to apply to **any** resource implemented as a MOOClet. These policies could be used for personalization and recommendation of Resources, or a wide range of reinforcement learning applications to adaptive experimentation.

Web Service API Calls for the MOOClet Framework

Name	Parameters*	Action
`assignVersionOfResource`	`learner_id, mooclet_id, policy, [policy_parameters]`	Assign version of MOOClet using current policy
`modifyVersion`	`version_id, mooclet_id, version_content`	Add or modify a new version for a MOOClet
`modifyVariable`	`learner_id, mooclet_id, variable, value`	Add or modify variable in Learner Data Store
`setPolicyAndParams`	`mooclet_id, policy, [policy_parameters]`	Change or update policy and parameters

*Parameters enclosed in [] are optional.

Figure 3.3: The key API endpoints for the web service that serves as the backend for MOOClets by providing Resource Versions to the Front-End Learner Interface, and allowing modification at any point via API calls to the Learner Data Store, Policy Set and Version Set

3.5 MOOClet Use Cases

In this section, we describe some use cases for the MOOClet framework through four illustrative examples - (1) *Motivational Messages*: Improving and personalizing motivational messages on edX, (2) *DynamicProblem*: enhancing online problems using a instructor-entered approach to experimentation, (3) *AXIS*: generating and experimenting with learner-sourced explanations, and (4) Personalized Problem Recommendation: recommending problems in a planetary-science MOOC on edX.

3.5.1 Use Case 1: Enabling Experimentation, Data-Driven Improvement, and Personalization of Motivational Messages

Instructors can encourage students by embedding motivational messages inside a tutorial page or quiz, such as before students attempt some problems or start an assignment. However, knowing which messages work best and for which students is not always easy in online learning environments. In this first example, we outline how we used our web-service to implement motivational messages on edX as MOOClets, and in doing so, enabled instructors to improve and personalize messages

given to learners. Then, we contrast our approach with existing independent systems that do not use a unified framework.

Step 1: Creating a MOOClet. Using the back-end admin panel of our web service, we create a new MOOClet instance called `MotivationalMessage`, and include an initial message Version to the Version Set via `modifyVersion` API call.

Version A: Learning can be challenging, every minute of effort moves you forward!

We link the MOOClet web service to the Resource edX using the `getVersion` API call via JavaScript embedded in an edX page to `MotivationalMessage` to obtain and place the message on the tutorial page.

Step 2: Choosing a Policy. We use the `setPolicyandParams` API call to assign `weighted_random` as the Policy for our MOOClet, with probability 100% (because it is the only possible version to be presented). Then, we use the `addVariable` API call to create a new `condition` variable in the Learner Data Store that will be automatically populated by the Policy with information on which Resource Version is assigned to each learner over time. Because we have a single message in the Version Set at this point, the set-up so far corresponds exactly to current practice. We now turn to how implementation as a MOOClet enables flexible future research.

Step 3: Experimenting with alternative Versions. To try an alternative message, we simply add it to the Version Set using another `modifyVersion` API call, and make an API call to `setPolicyandParams` to set the weights as [50%, 50%]. Then, we can compare the two versions using traditional A/B testing because the `weighted_random` setting for the Policy with its current parameters will evenly split the resources delivered to learners between the two versions.

Version B: Keep up the good work!

Step 4: Adding data to Learner Data Store. As a first step towards analyzing the data, we add information about the dependent or outcome variables to the Learner Data Store. In this case, we include two new variables, `motivational_message_rating` and `time_spent_on_problem`, to sit alongside the `condition` variable we added in Step 2. The `addVariable` API call allows flexibility in how such dependent variables are added. For example, we can use logging code in the Learner Interface to pull data from edX (or use data APIs is available), or download them into a spreadsheet. This flexibility allows our framework users to gather all the information necessary to analyze the experiments in the Learner Data Store.

Step 5A: Data-Driven Improvement. After running the A/B comparison with a sufficiently large group of students, in our case, say 150 people, we can look at the data and decide to change the probability of delivering a particular Resource Version using the `setPolicyandParams` API call. For example, we can switch from $[P(A) = 50\%, P(B) = 50\%]$ to $[P(A) = 20\%, P(B) = 80\%]$ if the `motivational_message_rating` or `time_spent_on_problem` are significantly higher for B than A. If the trend persists over time, we can eventually switch to $[P(A) = 0\%, P(B) = 100\%]$ as we become more confident that B is the better version among the two.

Step 5B: Personalization from data. We could instead choose to personalize the message delivered to students, such as sending Version A to students with a low grade in the course, and Version B to students with a higher grade. This can be done by adding the `course_grade` variable to the Learner Data Store, and updating the Policy from `weighted_random` to `weighted_personalization` using

the `setPolicyandParams` API call, and setting the Policy parameters to include some `IF-THEN` rules, which can be summarized by this pseudocode:

```
IF course_grade < 50%:  prob(A,B) = [100%, 0%]
IF course_grade >= 50%:  prob(A,B) = [0%,100%]
```

Notably, because the code or algorithm for assigning Versions is not stored in the Learner Interface, any `IF-THEN` rules can be defined for personalization, at any point in time, using any variables that can be added to the Learner Data Store.

Experimentation, Improvement, & Personalization without MOOClets. We contrast this with trying to achieve similar goals using standalone tools that do not use the unified MOOClet architecture. We focus on the very popular LMS (Learning Management System) Canvas to illustrate these challenges go beyond MOOC platforms, and because in our experience similar challenges arise in platforms that have siloed conceptualizations and implementations of software for experimentation, data-driven improvement, and personalization.

Experimentation. In Canvas, an independent LTI (Learning Tools Interoperability) tool was created [267] for doing randomized A/B comparisons. This tool only allows us to randomly divide students into different groups so that variations of course content can be presented to them. One drawback of this LTI tool is that the instructor is only allowed to edit the experiment before it starts. As soon as the experiment is started, the students will be distributed to the pre-allocated groups or conditions, and no changes, including adding a new condition or removing an ineffective one, can be made to improve the experiment– the entire experiment has to be removed.

Improvement. Transitioning from running experiments to making practical improvements to resources requires a longer timescale due to the overhead in switching between custom tools. To change which version is assigned, we have to remove the A/B testing tool, delete the alternative version, and revert back to regular Canvas. We have to choose between using standard Canvas to present one version, or embedding a new tool to randomly choose between multiple versions. Automated improvement of the kind shown in Use Cases 2 and 3 is certainly out of the question.

Personalization. Canvas has a separate tool that allows some Personalization, called MasteryPaths [217]. Using this tool, we could choose which modules to deliver to students based on what their accuracy was before. For example, based on the result of a pre-assessment quiz, we could put students into different groups so that each group has the same level of expertise in the subject. Then, we could design different paths for each group by delivering content webpage A vs content webpage B, tailored to the needs of the students of that group. MasteryPaths is a very natural way to do Personalization, but it is limited in only applying to a very specific set of data– the designer of the tool has to decide ahead of time exactly what variables might or might not be useful to personalize on. In a MOOClet, any variable stored in the learner data store can be used.

Another drawback is the very specific set of rules for personalizing content – using a particular graphical interface– whereas the MOOClet allows new Policies to be added, whether code, or even algorithms for problem-recommendation, such as those considered in Use Case 4. Finally, MasteryPaths cannot be used to do Experimentation – as mentioned that is a completely separate tool. That is problematic because it prevents us from testing whether our personalization approach is good, or discovering better ones. For example, we cannot use the same infrastructure/tool to randomize students to receive Content Type A vs Content Type B, and then obtain data about which is better for students with different levesl of knowledge.

3.5.2 Use Case 2: End-User Tools for Adaptive Experimentation

DynamicProblem [368] is an end user tool[3] that enables randomized experiments on the explanations, hints, feedback messages, and learning tips that show up after students attempt problems. This interface component (what is shown after submitting an answer) was implemented as the Resource component and linked to a MOOClet, and the Version Set contained explanations, hints, feedback messages, or learning tips as the Versions. [368] report three deployments using the DynamicRandomization Policy, which allows for machine-learning driven automated improvement using the Thompson Sampling multi-armed bandit algorithm. Student ratings of helpfulness of a given Version (e.g. hint/explanation/learning tip) was used as the outcome variable to be optimized for, and was logged and sent to the Learner Data Store. Instructors could also choose to use other MOOClet policies, and could pilot Versions using the WeightedRandomization policy, starting with equal probability of assignment and eventually moving to giving everyone a single Version by tweaking the weights.

DynamicProblem, provides a custom interface for instructors and researchers to author experiments and interact with the MOOClet backend, without using any API calls or programming. It also provides a Data and Policy Dashboard (Figure 3 in [368]) to allow instructors to see the behaviour of the system in real-time. This illustrates how the MOOClet architecture allows for the development of custom end-user tools built on top of its data structures and APIs, for purposes like authoring experiments, examining data, and interpreting algorithms for adaptive experimentation.

Although this functionality was not used in these deployments, the MOOClet would also have allowed Personalization of explanations/hints/learning tips as a function of any data about students that was added to the Learner Data Store, such as accuracy on previous problems, prior grades, information from a course survey.

3.5.3 Use Case 3: Learnersourcing Versions for Iterative Adaptive Experimentation

AXIS [367] (the Adaptive Explanation Improvement System) is a system that uses learner sourcing [163] to generate explanations and then adds these to an adaptive experimentation that eliminates lower rated explanations and keeps higher rated ones (as in Use Case 2). The original system was a one-off implementation using code written just for one purpose, making it a good candidate to re-implement using MOOClets to illustrate how its functionalities can be extended by the framework. The interface component displaying explanations was implemented as a MOOClet, and explanations generated by students that met a minimum length were added to the Version Set automatically, resulting in a series of iterative experiments with new Versions. The DynamicRandomization Policy was used with explanation rating as the outcome variable to optimize.

This MOOClet implementation provides several advantages. It allowed an interface to be built that let instructors review, edit, and/or remove student explanations (via API calls to the Version Set). Moreover, the outcome variable for optimizing explanations could be changed from ratings, to any other variable sent via API calls to the Learner Data Store, such as accuracy on a particular problem. As in Use Case 2, the MOOClet framework would also allow personalization of these

[3]It can be embedded into "any learning management system or MOOC platform that supports the ubiquitous Learning Tools Interoperability (LTI) standard" [368].

explanations, based on data like which explanations students have seen before, or their reading fluency on a survey.

3.5.4 Use Case 4: Personalized Problem Recommendation

In this final example, we use the framework to provide individualized problem recommendations to students in a planetary science MOOC on edX based on performance on prior problems. While the previous use cases focus on experimentation, this use case shows how a MOOClet can be used for personalized problem recommendation [200]. It also shows how our approach to designing software for dynamic experimentation and personalization generalizes beyond our specific web service instantiation, by having the Versions Set act as a proxy for serving content authored in edX, and going beyond the existing policies by using a policy external to the web service.

The Learner Data Store collects variables and sends them to an API service from an adaptive learning company, and which acts as a variant of Bayesian Knowledge Tracing [14].

We add resources to the Learner Interface using an LTI tool that allows us to embed various problem "windows" inside an edX MOOC. These windows show one problem at a time, and students click a "next" button to move on to a new problem.

There were four MOOClets in this context, one for each place in the course that an LTI problem window appeared. Each MOOClet had a Version Set of about 10 items, each item being a possible problem that the window could display. These problems were built inside the the edX course, and displayed within the LTI tool using the URLs. Each version in the version set was associated with its respective problem URL.

When learners attempt problems inside the content window, the application passes data from each problem to the Learner Data Store, recording the date and time, whether the attempt was correct, and the associated `problem_id`. Clicking the "next" button, triggers the MOOClet, which then uses its associated policy to select the next problem version to serve.

A notable feature about the Policy used in this case was that it involved an API call to an external web service. This external service was an API implementation of a variant of Bayesian Knowledge Tracing (BKT-Variant) [14]. For a given MOOClet and user id, the company's API would recommend one of the problems from the Version Set using BKT-Variant, which consumed variables from the Learner Data Store about that user's performance on past problems. The Policy was therefore realized through an API call to ExternalPolicy, with `user_id`, and `mooclet_id` as policy parameters. By linking `user_id` and `mooclet_id`, the Learner Data Store could receive a new entry every time a student viewed a problem in the MOOC. The Learner Data Store also received information about other MOOC courses that the student had enrolled in, and their course activity outside problems, such as the number of videos watched.

3.6 Discussion

Several insights and implications follow from our presentation of the MOOClet framework's motivation, design requirements, architecture, web-service implementation, and use cases.

Connecting the Architecture & Design Requirements. We highlight three reasons why the MOOClet framework separates the front-end Resource from the back-end databases (Version Set, Policy, and Learner Data Store) and provides APIs to access/modify these databases. This:

(1) Enables three typically siloed activities (Design Requirement D1, D2, D3) to share a common infrastructure, by simply changing the Policy (D6) for how alternative Versions are assigned: (a) Experimentation (Versions assigned with equally weighted randomization, e.g. 50/50); (b) Data-Driven Improvement (unequally weighted randomization) that is manual (humans choose weights) or automated (algorithms choose weights based on data about past students from Learner Data Store); (c) Personalization (Versions assigned based on data about current student from the Learner Data Store); (2) 'Future-proofs' infrastructure for research/practical activities that were not initially conceptualized but might be of higher quality, such as testing new Versions (D5), using new algorithms/decision-rules (D7), and using new data (D8). This allows agility in how research is conducted and speeds up iteration cycles, which is especially important when predictions of the most promising activities will change *after* a study/algorithm is deployed in the real-world; (3) Enables a common approach to be taken across multiple platforms (MOOClets have used been used in Canvas, edX, PCRS, Qualtrics, Mailservers, Twillio text messaging).

Where might readers apply the MOOClet framework?[4] Even without knowledge of a finalized experimental design or algorithm, which MOOClets could set the stage for future research using iterative experiments, dynamic data-driven improvement, and personalization? The potentially broad applications of the framework mean almost any component of an educational interface could be used. On the other hand, there are constraints in the kinds of work enabled by MOOClets (e.g. experimenting and changing assignment of alternative Versions based on past data). One consideration we use is which Resources are "sufficiently" modular/scoped for the goals of prospective users/stakeholders of MOOClet-enabled platforms. Since MOOClet-enabled work can clearly benefit from collaborators across disciplines (e.g. researchers in experimental psychology, learning analytics, reinforcement learning) and roles (e.g. researchers, instructors, instructional designers, programmers), we often have discussions with potential collaborators about candidates for potential MOOClets, with reference to a collaborator's goals, resources and constraints.

Such discussions led to the following applications, with over ten thousand learners. Course webpage components were implemented as MOOClets and used to vary Versions of Resources like: explanations, feedback on answers, motivational messages to solve problems, assignment to practice problems, instructions for how to engage with discussion forums, self-regulation activities that encourage planning, brief lessons teaching study strategies, and psychological interventions such as teaching a growth mindset. Implementing emails and text messages as MOOClets enabled testing of prompts for students to plan their work, reminders to start homework early, and activities for managing stress.

Implications for Educational Platforms, Developers, Instructional Teams, Researchers conducting experiments, Researchers publishing in data mining, reinforcement learning, and applied statistics. We hope the value to instructional teams of having course components implemented as MOOClets is clear. Even if one does not yet know the details, it leaves the option open for one's future self or collaborator to experiment with, improve, and/or personalize the course components in as-yet-unknown ways. Similar value accrues to educational platforms like LMS and MOOC providers, if they link (something like) the MOOClet architecture to particular course components. We have observed many platforms miss tremendous opportunities when they have developers implement a framework/tool for experimentation or personalization without con-

[4] Or a better version of MOOClets they are inspired to develop.

sulting the MOOClet architecture: Use Case 1 shows how Canvas has a tool for experimentation that enables no personalization, and vice versa not being able to randomize versions in a personalization tool and test the effectiveness of personalizing. The edX A/B testing tool and ASSISTments A/B testing tools cannot simply use WeightedRandomization to 'flick a switch' and transition from an experiment to giving the better version, they must replace the experiment.

Similar missed opportunities arise for any education researcher conducting a randomized experiment, whose code simply generates random numbers. Putting the thought into using something closer to a MOOClet allows the addition and removal of new conditions, making follow-up studies easier to run. Instructors have assurance there is a clear pathway to move from randomizing to giving a best version to students, and the potential to use reinforcement learning and bandit algorithms for automated improvement– and, uniquely, to *change* which outcome metric to optimize for without redeploying a webapp. Applied Statistics researchers can use MOOClets as sources of real-world data from adaptive experiments, and a unique opportunity to design and deploy adaptive experiments. Reinforcement learning and bandit researchers can use MOOClets as a test-bed for applying and evaluating algorithms for adaptive experimentation, with the rare capacity to choose actions, and get access to outcomes and contextual/state data in real-time. Data mining researchers can use MOOClets to evaluate a range of algorithms for personalization, individualization and content recommendation, in a real-world dynamic setting, as opposed to static 'found' data sets.

3.6.1 Limitations & Future Work

The MOOClet architecture aims to *enable* new activities, providing a "high ceiling" for what can be done, and making these activities easier than current practice. But it should be acknowledged that some setup and use of this expressive web service will require skills that not all potential users might have – such as using APIs – just as researchers and instructional teams may work with technical staff when doing field experiments or deploying algorithms. Future HCI need-finding work with stakeholders can explore the more usable/specialized end user tools, like graphical user interfaces built on top of the MOOClet data structures and APIs (like Use Case 2: End-User Tools for Adaptive Experimentation).

Relatedly, MOOClets lower barriers and can democratize access to valuable methodologies like conducting experiments and using bandit algorithms to put research into practice. However, this raises many questions for future work. How does one give users guidelines, training, and collaborative support in how, when, and why to use different methods? Under what circumstances are the potential benefits of greater access to new methods outweighed by the potential risks? For example, Use Case 2 and 3 using adaptive experimentation might increase the chances students get better explanations, but also increase the chances of bias or complications in statistical analysis of data from the adaptive experiment. The MOOClet framework doesn't claim to answer this question for a user, but to first make it possible to conduct real-world adaptive experiments and ask these questions, sparking discussions in education akin to the ones started many years earlier in industry settings. One way MOOClets could help in developing answers to questions about adaptive experimentation, is through facilitating collaboration with the applied statistics/biostatistics [254] and machine learning researchers developing methods to analyze such data. MOOClets can provide access to real-world data and a testbed for evaluating algorithms for adaptive experimentation.

Many challenges can arise in getting a front-end LMS (Learning Management System) to interact

with an external service like a MOOClet. Interoperability issues can make it hard to use an API call to a MOOClet's Version Set to display a Version in some LMSs, if they have restrictions on embedding code, or if a particular university prevents such use. Getting data from an LMS into a Learner Data Store can also pose technical issues – the LMS might not allow embedding of logging code, might not have APIs. Such data might then have to be manually downloaded, and then sent to the MOOClet. Beyond technical challenges, there are also security and privacy concerns in making student data more readily available (whether to benefit students or to help research) to both internal and external services. While the current web-service implementation uses de-identified data, there is always the risk of identification, and future work can explore how to enable dynamic improvement while integrating best practices for security and privacy.

Platform owners can also implement their own sand-boxed and internalized version of the open-source MOOClet web service, integrating into their platforms. They can also use the paper's conceptual insights about the architecture and design requirements to implement their own custom infrastructures. More broadly, we provide the MOOClet architecture and open source web-service so that future work can take inspiration, but then build better versions. MOOClets have the potential to provide inspiration for a range of cyberinfrastructure that supports the use of experimentation for dynamic improvement and personalization in real-world user interfaces.

3.7 Conclusion

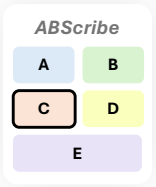
The AdapComp/MOOClet framework aims to transform components of educational interfaces into micro laboratories, by providing a software architecture and open-source web service to enable multi-disciplinary collaborations that improve student outcomes in tandem with conducting research. MOOClets enable experimental researchers and instructors to experiment iteratively and do the follow-up studies necessary to pin down what works and why, and to use data from experiments to more rapidly help future students. Statistics and machine learning researchers can investigate how to use algorithms for adaptive experimentation to automatically give students better resource versions, while collecting data that leads to statistically reliable conclusions. Data mining researchers can apply different algorithms for personalization and use data from real-world courses to improve them, while giving better and better problem recommendations to students. The MOOClet architecture's 'future-proofing' allows researchers to be more agile in incorporating novel ideas from unanticipated collaborators, adapting algorithms to the messiness of real courses, and using constantly emerging real-world data. Going beyond the use cases presented, how can we each leverage MOOClets for iterative experiments that lead to data-driven enhancement and personalization of educational resources, in real-world online courses?

3.8 Acknowledgements

We thank members of the Intelligent Adaptive Interventions lab for their input on the paper, and Sam Maldonado for his work on the MOOClet web-service implementation. This work was supported by the Office of Naval Research (ONR) (#N00014-18-1-2755) and the Natural Sciences and Engineering Research Council of Canada (NSERC) (#RGPIN-2019-06968).

Chapter 4

The ABScribe Interface



> The only kind of writing is rewriting.
>
> *Ernest Hemingway [130]*

Research Context: The framework introduced in the previous chapter solves the challenge of deploying continuous experiments involving multiple content variations packaged into AdapComps/MOOClets. However, the current authoring interfaces used to create those variations lack native support for parallel creation and management of those variations without overwhelming users. In this chapter, we describe interface elements that extend a workflow as intuitive as editing a document in a word processor to make parallel exploration of multiple content variations much easier, using generative AI to accelerate this process. This chapter addresses the third primary research question.

> *Q3: How might we support content authors in efficiently exploring and organizing multiple alternatives without overwhelming them, and how might generative AI assist in this process?*

The answer is to augment a familiar document editing interface with tools for parallel editing across multiple variations. In ABScribe, variation fields can be linked to the corresponding AdapComps described in the previous chapter, with the Variation Accordion providing an intuitive front-end for navigating the AdapComp Version Set.

Related Publication:

Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. *ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models.* In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1042, 1–18. https://doi.org/10.1145/3613904.3641899

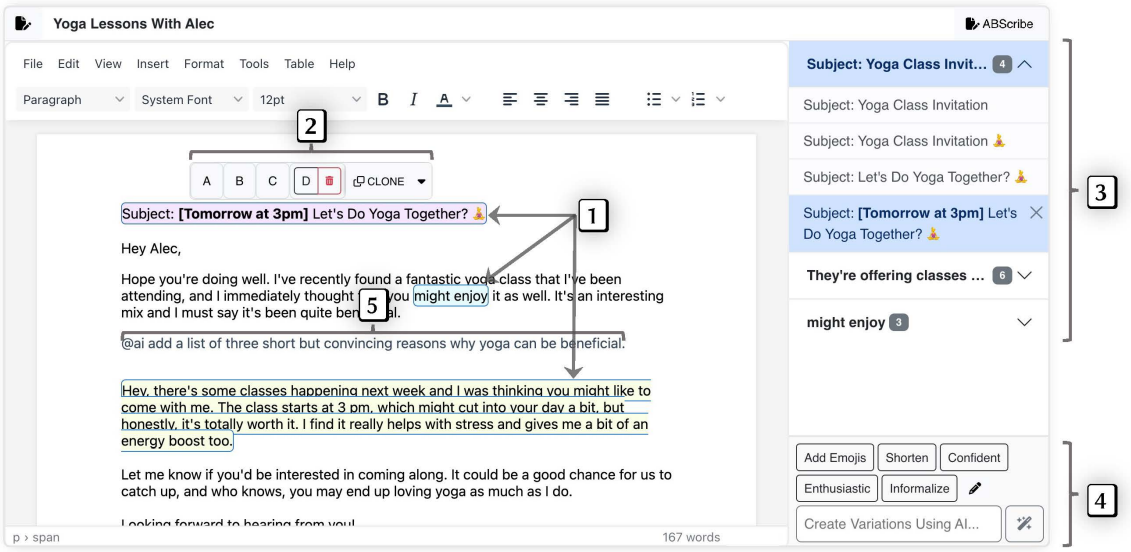


Figure 4.1: The ABScribe Interface: (1) **Variation Components**: Multiple variations are stored within text-segments that do not break the flow of the draft. (2) **Hover Buttons**: Users can swiftly compare multiple variations by hovering over buttons placed above the selected Variation Component, or clone and edit them in-place. (3) **Variation Accordion:** Users can view multiple variations and navigate through them using an organized accordion structure. (4) **AI Buttons**: Users can quickly create variations using AI by typing instructions auto-converted into reusable buttons that can be applied to other Variation Components. (5) **AI Insert**: Users can insert text from GPT-4 directly into the document by typing '@ai <prompt>' and pressing enter.

Abstract: Exploring alternative ideas by rewriting text is integral to the writing process. State-of-the-art large language models (LLMs) can simplify writing variation generation. However, current interfaces pose challenges for simultaneous consideration of multiple variations: creating new versions without overwriting text can be difficult, and pasting them sequentially can clutter documents, increasing workload and disrupting writers' flow. To tackle this, we present ABScribe, an interface that supports rapid, yet visually structured, exploration of writing variations in human-AI co-writing tasks. With ABScribe, users can swiftly produce multiple variations using LLM prompts, which are auto-converted into reusable buttons. Variations are stored adjacently within text segments for rapid in-place comparisons using mouse-over interactions on a context toolbar. Our user study with 12 writers shows that ABScribe significantly reduces task workload ($d = 1.20, p < 0.001$), enhances user perceptions of the revision process ($d = 2.41, p < 0.001$) compared to a popular baseline workflow, and provides insights into how writers explore variations using LLMs.

4.1 Introduction

Revision is an essential part of the writing process [387, 130, 106, 223]. Professional writers often write and rewrite text hundreds of times [325, 53] and recommend rewriting as a core strategy for writing well [387, 53]. Effective revision goes beyond minor editorial changes, and may help writers rework ideas, and powerfully affect their knowledge [91, 325] as they explore alternative variations to find a line of argument [91]. The revision process is *iterative* [82, 91]: happens in repeated cycles, throughout the writing process, *granular* [301, 223]: happens at the word, sentence, or paragraph-level, and *non-linear*[54, 325]: requires constant reconsideration of potential variations of existing text elements throughout the passage. Current writing interfaces tend not to support a non-linear revision process and predominantly support linear representations of revision history (e.g. revision history in popular word processing software such as Google docs or Microsoft Word). While these tools do support iterative and granular edits, it remains difficult for writers to *simultaneously* consider *multiple* writing variations and to organize them without replacing earlier text, or cluttering documents when writers resort to pasting them in sequence. Insights from HCI and traditional design practice suggest that simultaneous consideration of multiple (at least 5) variations can lead to better ideas [345] and avoid fixation [151, 345, 112, 42]. We hypothesize that this may apply to writing, provided that writers are given adequate support in managing multiple variations with minimal workload.

Advanced Large Language Models (LLMs) such as ChatGPT [140], GPT-4 [249], PaLM 2 [255], and LLaMA 2 [347], can enable writers to generate multiple variations of text via prompting [37, 69], potentially reducing the workload of generating text variations. However, easier generation can exacerbate challenges surrounding the systematic storage, comparison, and modification of multiple variations with existing chat-based and in-place editing interfaces where users are required to find text variations in linear chat histories or store them in their document editor as comments or separate in-line text blocks. As text variations become easier to generate using AI, they become harder to manage.

Recent HCI studies on LLM-based tool design have mainly focused on prompt engineering [31, 380, 104] and exploring the generative capabilities of LLMs [378, 191, 372, 214]. For example, Zamfirescu-Pereira et al. investigated ways to support non-AI experts with crafting effective prompts, and Yuan et al. [378] explored how users might use LLMs for creative writing. They found that the output of the model did not need to be perfect to be useful to users. Many users found the output useful, even if they had to significantly revise the text or chose not to incorporate it into their final draft [378]. This highlights the potential value of designing affordances that help manage *imperfect* AI-generated variations. Even if these variations don't make it into the final text, they might still be valuable to consider.

In this paper, we present ABScribe[1]–a novel writing interface that supports the rapid exploration of multiple writing variations in LLM-based human-AI co-writing tasks. We draw inspiration from Kim et al. design framework which shows the potential for object-oriented interactions with LLMs to encourage iteration and experimentation during writing [164], and propose a suite of five interface elements that support writers in swiftly exploring multiple writing variations by interacting with an ensemble of five interface elements: (i) **Variation Components** that store multiple human and

[1]We name our system *ABScribe* to reference how we label multiple variations using the alphabet. Note that we support variations beyond just A and B. The Hover buttons can include multiple variations: A, B, C, D, E, etc.

AI-generated variations within flexible text segments in a non-linear manner, without overwriting text; (ii) **Hover Buttons** that reveal corresponding versions inside a Variation Component when users hover their mouse over them, allowing for rapid comparisons without breaking text flow; (iii) the **Variation Accordion** that organizes all variations in a navigable format; (iv) **AI Buttons** that automatically encapsulates LLM instructions into reusable buttons that can be applied across different Variation Components; and (v) **AI Insert** that allows writers to insert LLM-generated text directly into the document (see Figure 5.1).

To validate our design, we conducted a controlled evaluation study and interviews with 12 writers comparing ABScribe with a widely-used baseline worklow consisting of an AI integrated rich text editor based on GPT-4, with a chat-based AI assistant. Our findings demonstrate that ABScribe significantly reduces subjective task workload ($d = 1.20, p < 0.001,$), and enhances user perceptions of the revision process ($d = 2.41, p < 0.001$), compared to the baseline. The key contributions of our work are as follows:

1. The design and implementation of ABScribe, an LLM-enhanced writing interface that supports the rapid exploration of multiple text variations in human-AI co-writing tasks.

2. The results of a 12-participant user study with writers demonstrating the efficacy of the ABScribe interface ensemble and its advantages over a commonly used baseline workflow, and user perspectives on how writers explore multiple variations in human-AI co-writing tasks using a linear and non-linear revision process.

4.2 Related Work

We review literature in HCI and traditional design practices on weighing multiple alternatives, and discuss the relevance of this design method when revising writing, guided by they theory on revision process in writing. We delve into difficulties that arise when trying to support multiple variation exploration using existing editing interfaces and contrast between *chat-based* and *in-place* interfaces to situate our design within a broader class of Human-AI writing interfaces.

4.2.1 Exploring Multiple Variations

HCI and traditional design practice encourages the parallel exploration of multiple variations to help avoid fixation on a singular idea [151, 108], to reduce the chances of eliminating rough but innovative ideas due to premature evaluation [112, 42], and to make us less prone to inflated subjective appraisals by giving us an opportunity to critically assessing ideas in relation to each other [345, 42]. In this paper, we hypothesize that such parallel exploration of multiple variations may apply to the revision process during writing. Much like how a naive, linear implementation of an iterative design approach encourages the sequential refinement of ideas, when writers don't have a way to organize and work with multiple text variations, they may end up committing to ideas too early, and focusing too much on surface level edits to refine their draft.

This is problematic because when we turn to research on the revision and the writing process, we see that revision goes beyond surface level edits [91], encompassing deeper writing subprocesses such as revising and evaluating ideas [92] and meaning discovery [325]. Experienced writers treat revision as a recursive, non-linear process [325, 301], and engage with the text in repeated cycles,

with multiple objectives including finding the form or shape of an argument [325], experimenting with vocabulary and style [137], and going back and forth between multiple composing activities as writers revise text [87].

In addition to the rich-body of work underscoring the important and complex role of revision in writing, researchers have explored the benefits of adopting design language in writing pedagogy, such as characterizing writing pedagogy as a wicked [286] design thinking problem [196, 271]. There has also been some valuable work in HCI to support novel editing practices, such as supporting constraints and consistency in maintaining domain-specific terms across complex documents [116] using persistent, reified [20] text selections, and present the idea and implementation of a *varientlet*, that allows writers to store and compare two variations. However, further innovation in this space is needed to design affordances for writers that support the simultaneous consideration of multiple variations during the revision process.

In this paper, we contribute to this line of work, and present a suite of interface elements that work together in supporting writers with the rapid exploration of multiple text variations in a non-linear fashion, in alignment with the nature of the revision process, and offer empirical insights into the applicability of design ideas in HCI on the parallel consideration of variations to the specific task of revision in writing.

4.2.2 Working with Multiple Variations from Large Language Models

As we work toward leveraging advanced LLMs such as ChatGPT [140], GPT-4 [249], PaLM 2 [255], and LLaMA 2 [347], which can enable writers to generate multiple variations of text based on different parts of their writing using prompts, HCI researchers looking to design writing interfaces are faced with several challenges. These include dealing with the non-deterministic nature of these models [164, 152], systematically exploring their capabilities [378, 191], and making prompt-writing easier for AI-novices [380, 305].

There has also been work on managing the output from generative AI to support the exploration of variations in different contexts such as exploring images [31, 190], and multi-modal interactions beyond text prompts to explore generative AI [207]. However, our understanding of how to best organize the prolific output in AI-augmented writing workflows is still limited.

We draw inspiration from Kim et al's work on the use of object-oriented interactions and reification [117] to encourage writers' experimentation of LLM output, as well as prior work on revision control in writing [116, 115] and other domains [128], and offer the design and implementation of a novel interface that tackles the problem of how to effectively *organize* multiple variations from an LLM in a way that minimizes task workload while supporting parallel exploration.

4.2.3 Chat-Based and In-Place Human-AI Co-Writing Interfaces

To help ground our interface design, we distinguish between two types of Human-AI Co-Writing interfaces into two types: conversational interfaces such as ChatGPT and Bard, and In-Place interfaces that directly inserts or modifies text in the document.

Chat-Based Interfaces: Currently the dominant mode, chat-based interfaces, like ChatGPT [140], Bing Chat [280], and Bard [**bard**], have gained immense popularity. These conversational interfaces are highly intuitive, and mimic human-to-human chat interactions, but lack scaffolding for crafting

prompts, which can be difficult for novice AI users [380, 183]. Another significant limitation is the linear chat-log structure. In contrast to the non-linear nature of how revision happens in writing [325, 91] the text-variations generated using a chat-based interface are buried within linear chat-logs, impeding parallel exploration of multiple variations in-place, where the writer is editing the document.

In-Place Editing Interfaces: This type offers closer integration between the human and AI writer during the text editing process by adopting a more *What You See Is What You Get* (WYSIWYG) [27] approach where AI-generated text modifies the human text and vice versa. This offers increased flexibility over the edited content compared to chat and form-based interfaces by allowing users to edit individual sections of the text.

Recent research prototypes for LLM writing tools such as Wordcraft [378] and CoAuthor [191] allow for in-place editing. Wordcraft [378] melds an in-text interface with diffrent options for users to continue a narrative based on prior text and replacing text selections with AI-modified content. CoAuthor explores GPT-3's capabilities by capturing deep interactions between writers and GPT-3 via a similar in-place editing interface where users receive multiple edit suggestions from the model. Commercial tools like Grammarly[2] and Wordtune[3] also allow users to enhance their writing by revising text using AI-driven suggestions. However, once generated by the AI and modified by the user, previous text can become obscured or lost as older text is superseded by new edits. Even if text is auto-saved, it is often preserved as linear version histories, as is the case in Google Docs (an online word processor), or undo/redo histories, making it difficult to work with multiple text variations concurrently.

In our design, we adopt an in-place editing interface in a GPT-4 powered research prototype, offering a solution to overcome challenges surrounding the management of multiple text variations in human-AI co-writing tasks. We carefully construct a baseline interface that represents current workflows, providing fresh empirical insights based on our interviews with writers. These insights help us understand user perceptions of the revision process and explore how differences between in-place editing and chat-based AI writing companions impact their workflow.

4.3 Designing ABScribe

In this section, we describe the design requirements for ABScribe and the interface elements that we developed to address those requirements.

4.3.1 Design Requirements

We surveyed literature on several key areas critical to our goal of facilitating the swift exploration of multiple writing variations using LLMs: the role and nuances of revision within the writing process [325, 91, 212, 123, 306, 93]; HCI design philosophies that emphasize the consideration of multiple ideas before evaluation [345, 112, 79]; principles on reification and reuse for designing visual interfaces [20, 117]; and the latest research on utilizing LLMs in writing interfaces to foster experimentation and creativity [378, 164, 191]. Based on this, we formulated an initial set of design requirements, which are summarized below.

[2] grammarly.com
[3] wordtune.com

Requirement 1: Minimizing Task Workload while Exploring Multiple Variations of Text Drawing from HCI and traditional design principles, we emphasize the importance of exploring multiple ideas concurrently. Instead of refining a single solution to "get the design right", these disciplines encourage the iteration and evaluation of multiple solutions in parallel to ultimately "get the right design" [345]. We hypothesized that parallel exploration of text segments could aid writers during the revision process. We also considered that during the writing process writers frequently struggle with cognitive overload [221] and that even small demands on working memory can lead to decreased fluency [276]. With this in mind, we hypothesized that an increase in writing variations could worsen this. As such, we aimed to design an editing interface that provides affordances for creating and comparing multiple writing variations without overwhelming the user.

Requirement 2: Support visually-structured management of variations Documents can become quickly cluttered when trying to explore multiple variations of different text segments using current editing interfaces. Furthermore, the potential for LLMs to enhance writers' ability to generate and revise multiple parallel variations of text further exacerbates the issue of clutter. Together, these factors highlight the need for a visually structured approach to manage variations. Our goal was to support writers in seamlessly integrating LLM-generated variations without cluttering the document or erasing existing content to retain the ability to simultaneously consider multiple variations. By presenting users with a range of variations, we give them the opportunity to select their favorites while simultaneously discarding less-favored alternatives [112].

Requirement 3: Support context-sensitive variation comparison and revision In a linear document editing interface, we found it difficult to maintain a sense of the surrounding text to situate new variations within existing context. This was particularly poignant when creating and comparing variations for smaller and embedded text segments - eg. A sentence mid-paragraph or paragraph mid-section - which disrupted the text flow. Maintaining text flow is crucial since writers need to engage with information processing tasks such as ensuring the document maintains cohesion which requires matching to surrounding text [220] This becomes increasingly challenging as the document holds more and more variations of text segments. Our objective was to design an interface that allows writers to systematically evaluate these variations within context, eliminate less-favored options, and generate new iterations based on existing ones.

Requirement 4: Supports revision-centric, reusable, and non-linear LLM usage Recognizing that revision is inherently nonlinear— with writers often revisiting earlier sections of a passage— and recursive, manifesting in repeated cycles throughout the writing process [325, 91], we aimed to align our LLM integrations with this fluid, iterative nature of revision. Our goal is for writers to be able to use LLMs to manipulate text segments of varying lengths and refine them as needed in a way that is natural to their non-linear and recursive process. To enable this, we draw inspiration from design principles for visual interfaces, focusing on reuse, polymorphism, and reification [20], and regard LLM prompts as reusable, polymorphic commands that can be applied to targeted text segments of varying lengths, transforming them into first-class objects. Recent research on designing LLM-powered writing interfaces has highlighted the value of viewing components of the LLM generation pipeline as interactive objects in supporting iteration and experimentation [164]. We adopt aspects of this approach in our design, such as reifying [20] LLM prompts into reusable AI buttons, and turning text segments into interactive Variation Component objects or *cells* [164].

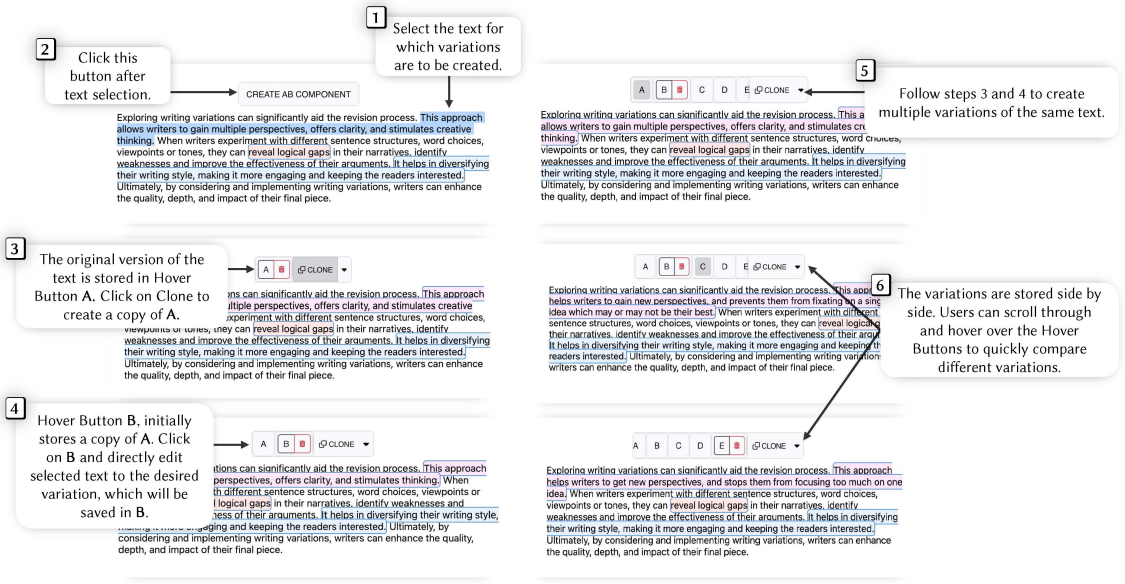


Figure 4.2: **Hover Buttons & Variation Components:** ABScribe supports the ability to store multiple writing variations in a variation component. These variations can be easily compared and swapped as shown above.

4.3.2 Interface Elements

We addressed these four design requirements by developing five interface elements using an iterative design process. To get the right design [345] the lead author iterated through multiple versions of each interface element, and tested them with a total of five pilot users during in-depth brainstorming and design-review sessions over six months.

Variation Components: Users can select any part of the text and create interactive Variation Components that can hold multiple writing variations (Figure A.1, Part 1). Newer variations can be added to an AB Component without overwriting existing variations (Figure A.1 Parts 2 and 3), and without breaking the flow of the passage.

Hover Buttons: Variations are represented using AB Hover Buttons that are dynamically placed above the active AB Component. Users can hover over each button to reveal the corresponding variation in the context of the surrounding passage (Figure A.1, Parts 4 and 5). Moving the cursor away from the AB Hover Button reverts the AB Component back to the selected variation, allowing users to quickly compare between the selected and the hovered variation. Users can click on the Hover Button to select the variation and edit them in place, or discard it by clicking on the trash icon (Figure A.1, Part 6).

Variation Accordion: To help writers view multiple variations together, and navigate through them more easily, we pair the Hover Buttons with an accordion structure where each Variation Component has its own header, and all corresponding variations are stored underneath. Clicking on the variations in the accordion dynamically re-positions the Hover Buttons above the corresponding Variation Component, and vice versa, allowing users to manage multiple variations in a visually structured manner (Figure 4.3)

The Variation Components, Hover Buttons, and the Variation Accordion work together to ad-

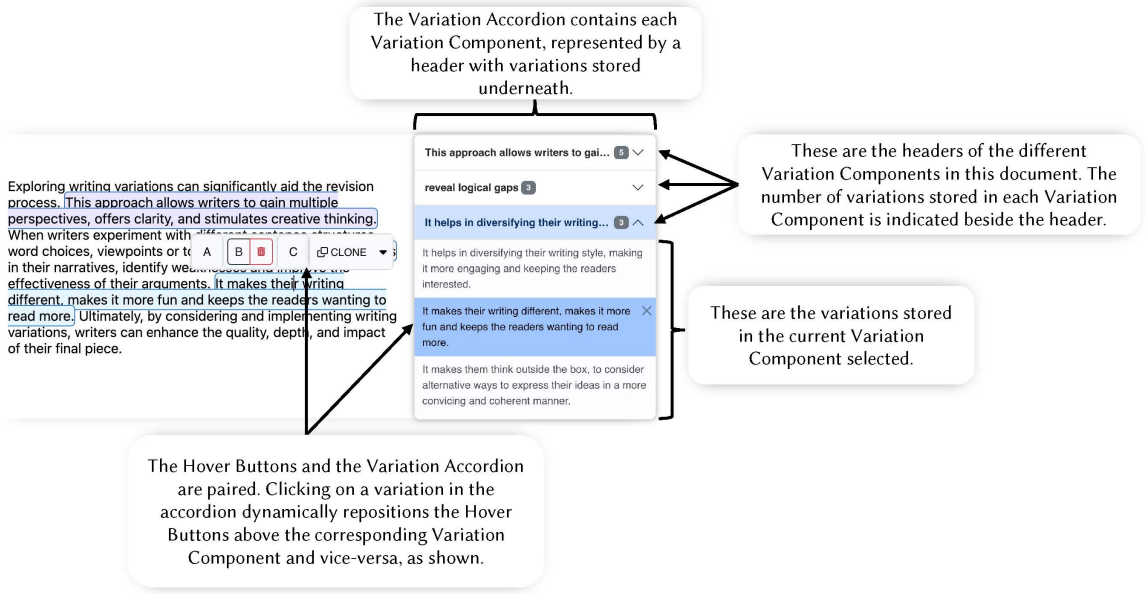


Figure 4.3: **Variation Accordion:** The Variation Accordion is an alternative method to viewing existing variations and is especially useful in viewing multiple variations side by side

dress R1, R2 and R3. To tackle R4, we developed two interface elements.

AI Buttons: Users can generate new variations by selecting an AB Component and typing instructions to the AI (Figure 3, Part 1). Instructions are automatically converted into labeled buttons (Figure 3, Part 2). The labels are generated using the LLM. As users experiment with newer variations using the AI, they create a set of custom AI Variation Buttons that they can reuse to apply to different parts of the passage, making these buttons reusable (Figure 3, Part 4). The prompts and labels for the buttons can be edited and improved over time. (Figure 4, Part 5). This allows writers to not only create a set of variations for a particular AB Component, but also design a set of buttons reflecting the kinds of variations they might want to generate for other parts of the text in the future, akin to a custom Swiss army knife for variations.

AI Insert: Users can insert text from the AI model anywhere within the passage by writing instructions to the LLM in the following format: @ai <prompt>. The text is generated and shown to the user in real-time, and they have the option to accept or discard the AI generated output, as well as revise the prompt to regenerate the output if it doesn't match what the user is looking for (Figure 4.5).

4.4 Evaluating ABScribe

To validate our design, we conducted an within-subjects evaluation study where we compared ABScribe to a carefully constructed baseline interface. Refer to Appendix A.2 to see a screenshot. The Baseline interface featured rich-text editing capabilities commonly found in word processing software such as Google Docs and Microsoft Word, as well as a conversational AI assistant similar to ChatGPT, and the ability to incorporate AI generated text into the document without the need to copy and paste to represent the tighter AI integration available in modern AI editors such as

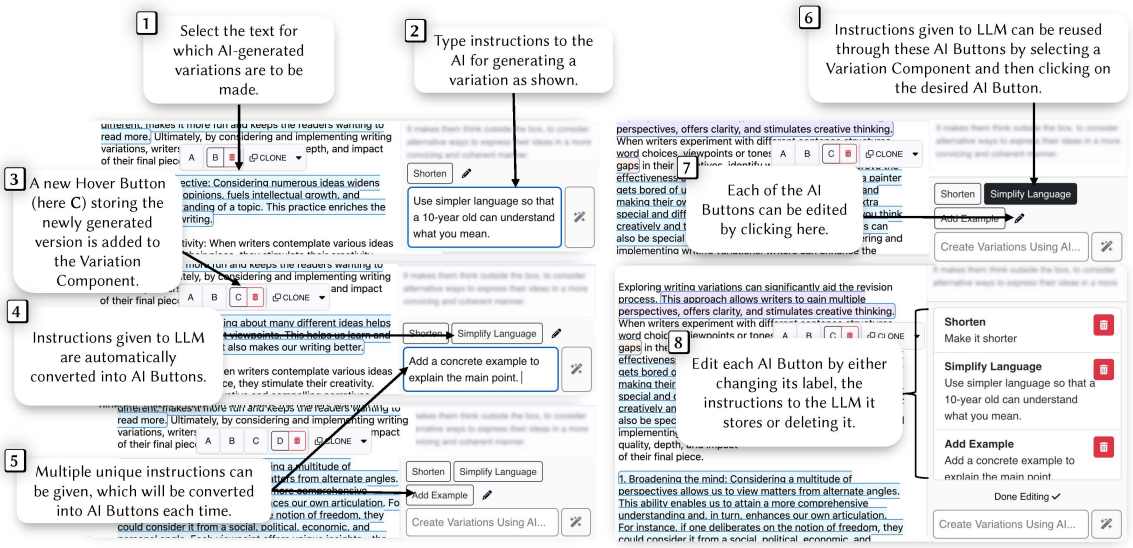


Figure 4.4: **AI Buttons:** Variation Components can also be edited using the AI buttons, which lets users specify alterations for an chunk. Descriptive labels are automatically generated for each AI button and each button can be reused and edited.

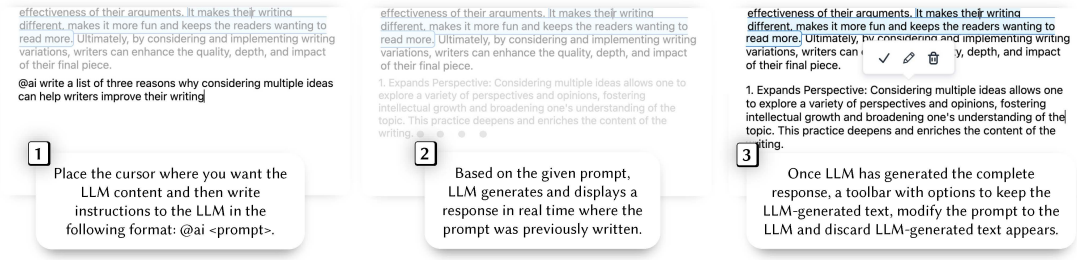


Figure 4.5: **AI Insert:** The AI Insert feature provides the ability insert LLM-generated text directly into the document, providing tighter integration between the Human and AI generated writing workflow. Users can see the AI generated content in real-time and choose insert or delete the output, or revise the prompt, giving users more control over what is included in their document.

Notion.AI[4]. To minimize potential confounding factors due to tangential differences between the study conditions, we maintained the same overall layout for the common UI elements such as the width of the sidebar, placement and dimensions of buttons and text, font size, and color, and used the same underlying LLM model, GPT-4, to implement the generative AI features. We sought to answer the following research questions:

RQ1: How does ABScribe influence **user perceptions of the revision process** for AI-assisted exploration of writing variations when compared to the AI-integrated Baseline interface?

RQ2: How does ABScribe influence **subjective task workload** for AI-assisted exploration of writing variations during text revision when compared to the AI-integrated Baseline interface?

4.4.1 Participants

We recruited 12 writers (5 women, 7 men), aged 18 to 34, all of whom reported proficiency in reading and writing in English. They were screened for prior experience in a broad variety of both fiction and non-fiction writing genres. We included a range of prior experience levels of AI tool usage. This was done to ensure that our findings were not tied to specific genres or specific AI usage traits. Moreover, all participants had sufficient writing experience to be able to comment on the revision process and its applicability to different kinds of writing. See Table 5.1 for writer profiles.

4.4.2 Tasks

Each participant engaged in two guided writing tasks which were randomly paired with the two counter-balanced study conditions. Our choice of tasks—writing an email and a social media post—aimed to provide an ecologically valid writing experience that fits the timing constraints of our study and served as realistic use-cases for LLMs, in alignment with recent HCI studies on human-AI co-writing [164, 110, 41], and commonly advertised applications of commercial AI-assisted writing tools, such as Grammarly Go, Respondable, and Copy.ai. We selected scenarios by considering situations that: (i) would be easy for users to imagine, such as emailing a professor or seeking a job as a copywriter; (ii) offer opportunities for exploring variations, such as devising multiple alternatives for a subject line or altering a sentence to maximize reader engagement. We wrote and tested the prompts to ensure the output was reasonable and of consistent length. For example, after initial rounds of testing, we found that the output generated by the AI was too long, and so we added 'Keep it within three paragraphs' to both prompts. We asked participants to generate a draft using the same prompts for all scenarios. This approach was designed to keep their focus on exploring and organizing variations, rather than engaging in prompt engineering on the initial draft. It also helped maintain relative consistency in the generated output across participants and study conditions. Refer to Appendix A.1 for additional details on the task scenario descriptions and prompts.

Then, with either ABScribe or the Baseline interface, participants were asked to use AI to explore eight variations (increasing or decreasing length, formality, word diversity, adding emojis, and two variations of their choosing) of three distinct text segments (title/subjectline, third sentence of the

[4] notion.ai

second paragraph, entire third paragraph), summing up to twenty-four variations. This approach ensured variation exploration of consistent number, size and variety across study conditions, while affording some scope for creativity as, for two out of the eight variations for each segment, participants had the autonomy to craft variations based on their preferences. We selected the variations by considering potential modifications writers might want to apply to their drafts, including changes in length, tone, and word choice. We then tested these modifications to ensure their feasibility with both ABScribe and the baseline interface.

4.4.3 Measures

To assess subjective task-workload, we used the widely used NASA-TLX [124] procedure with weighting. To quantify the specific aspects of the LLM-assisted revision process that we aimed to improve, we also asked participants to rate their agreement on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree), similar to prior work [194, 164, 191]:

1. **Variation Granularity:** *I felt like I could work with multiple (more than 5) writing variations of different fine grained parts of the text (e.g. word, sentence, paragraph) using this tool.*
2. **Variation Search:** *I felt like after creating all these variations, I could find previous variations when I needed them (e.g. when trying to create a new variation based on an existing variation I created earlier in the writing process).*
3. **Prompt Reuse:** *I felt like after creating all these variations, I could reuse my previous instructions or prompts to the LLM without having to rewrite them often.*
4. **Variation Comparison:** *I felt like I could identify fine-grained differences between multiple variations using this tool.*
5. **Variation Editing:** *I felt like I could systematically edit new variations without losing existing variations or cluttering the document using this tool.*
6. **Variation Control:** *I felt like I had control over which variations I wanted to keep, discard or change.*
7. **Variation Divergence:** *I felt like exploring multiple variations using this workflow will help me come up with variations that are surprisingly different.*
8. **Draft Quality:** *I felt like exploring multiple variations using this workflow will help me have better final draft.*
9. **Intent Match:***I felt like exploring multiple variations using this workflow will help me come up with variations that are closer to what I want to say.*
10. **Variation Diversity:** *I felt like I could create variations with a lot of variability in word choice, style, and tone of voice using this tool.*
11. **Document Clutter:** *I felt like after creating all these variations, the document became cluttered.*

4.4.4 Procedure

Participants began by signing a consent form and completing a survey that captured demographic data, prior writing experience, and familiarity with AI-assisted writing tools. Conducted via video-conference, the entire study lasted approximately 1.5 hours. Participants accessed the prototypes through their web browsers, mirroring how they typically access popular AI editing tools like ChatGPT and Grammarly. Conducting the study online enabled us to engage a diverse group of writers beyond Canada. After a brief introduction outlining the study's objective—to investigate various writing variations using LLM-integrated editing tools—participants undertook two 15-minute tasks. Before starting each task, we demonstrated how the tools functioned and gave participants an opportunity to try them, ensuring they felt comfortable. After each task, participants completed the NASA-TLX and 11 Likert-scale measures. These measures offered insights into the writers' perceptions of the revision process and prompted them to reflect on specific aspects of the revision process that we seek to improve through our design. The evaluation concluded with a recorded 30-minute semi-structured user interview on their experience with each interface. As a token of our appreciation for their participation, each participant received 30 Canadian dollars.

4.4.5 Analysis

Our data comprised interview transcripts, task observation notes, and the NASA-TLX and Likert-scale ratings for each condition. We coded and analyzed the interview transcripts and task observation notes using reflexive thematic analysis [33] through an inductive-deductive lens. The theory on revision-focused exploration of writing variations served as a pre-existing code guiding our interpretations.

In a within-subject design, we use pairwise one-sided t-tests to compare *sum* of scores of NASA-TLX and our Likert-scale measures on the revision process. T-test was shown to be robust for aggregated data of this kind. [369]. Additionally, we aim to check for normality. We aimed to determine if ABScribe presented significant improvements over the baseline, prompting us to select a one-tailed test with hypothesis B < A for task workload and B > A for level of agreement on the efficacy of the revision process.

We performed an apriori power analysis for a pairwise one-sided t-test, showing that we can detect with 80% power at least $d = 0.8$ effect size with sample size $N = 12$ participants for a significance level $\alpha = 0.05$.

4.5 Results

> *"The user interface for A [ABScribe] made comparisons, easy storage and access of those variations much easier than B [Baseline]. The fact that after you wrote a prompt, it instantly assigned a button to it that you could access later, was incredibly useful. It made it such that you could actually play with a more precise number of variations than I previously could and the fact that you could manually edit them and then again, quickly, have a way to play with the variation made it much more practical as a writing tool, there was a lot less physical effort involved to streamline that process. It was a wonderful, very smart way of dealing with the problem of clutter on the page." – W1*

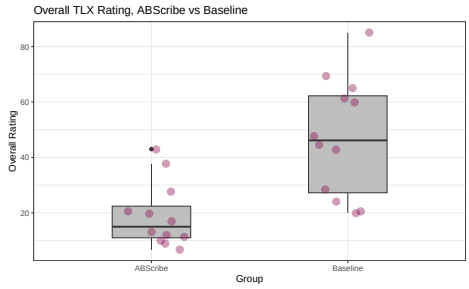


Figure 4.6: Overall results on NASA TLX subjective task workload

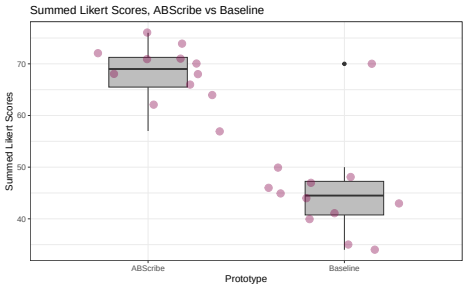
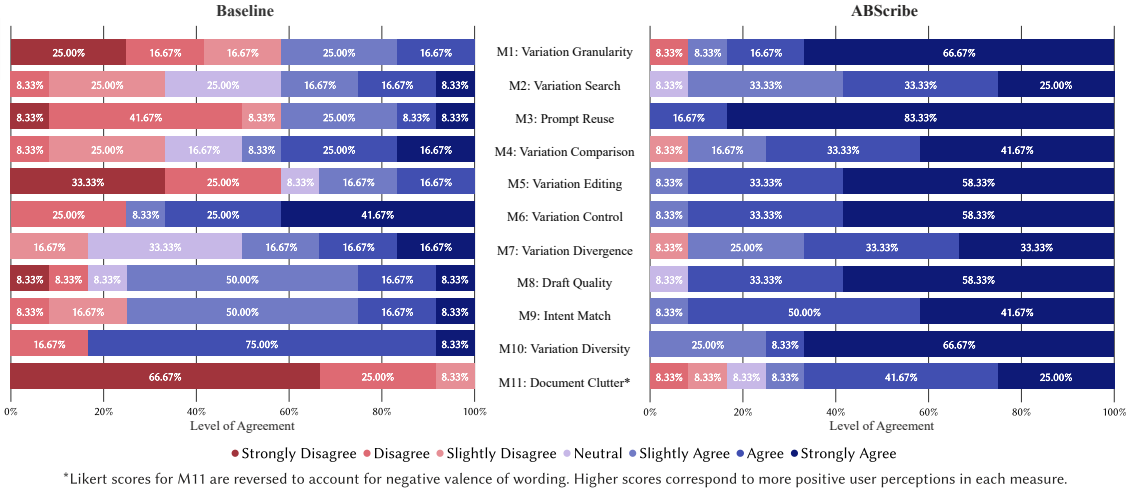


Figure 4.7: Summed Likert ratings for users' perceptions of the revision process



*Likert scores for M11 are reversed to account for negative valence of wording. Higher scores correspond to more positive user perceptions in each measure.

Figure 4.8: Responses to Likert-Scale Measure on the Revision Process for Exploring Multiple Writing Variations. Higher the agreement level, the more positive the user perception.

The overall response to ABScribe, as exemplfied by by W1's comment and Figures 4.6, 4.7 and 4.8, was positive, with a significant increase ($d = 2.41, p < 0.001$) in the summed agreement levels on the efficacy of the revision process (RQ1), and a significant reduction ($d = 1.20, p < 0.001$) in TLX rating for subjective task workload (RQ2) when compared to the Baseline interface.

To gain deeper insights into the factors behind the reduction in task workload and the increase in user perceptions, we conducted a reflexive thematic analysis on the semi-structured user interviews. *User perceptions on the revision process* groups findings relating to how the users perceived changes to the process and outcome of creating and managing variations as well as their interactions with the LLM. *Subjective task workload* groups findings relating to the ease of specific tasks during the writing and revision process as well the ease of specific interactions with the interface to accomplish those tasks.

4.5.1 RQ1: User Perceptions on the AI-Assisted Revision Process

F1: ABScribe lessens pressure to commit early to an initial idea and nudges users to explore a greater quantity of variations than the Baseline workflow. The non-linear approach to storing multiple variations within AB Text Components without cluttering the document,

and the ability to switch between variations using the Hover Buttons and the Variation Accordion made some participants feel less pressured to commit to an initial variation before considering multiple options. For example, W3 noted: *"I would probably feel more pressured to just kind of work on one sentence and come up with a couple of variations and change it immediately. I would feel like I have to commit pretty early on, rather than generating a number of variations, trying a bunch of different tracks and sort of different timelines, seeing how each of them turns out and performing a master comparison at the end. So I think it had a significant effect, or would have a significant effect on my behavior, certainly doing the same task for both conditions where I was trying to deal with a bunch of different versions and trying to change them and revise them differently. It was vastly more difficult in B [Baseline] and a nonlinear approach seems to make much more sense."*

W2,3 and 12 specifically commented that the lack of interface clutter due to the way AB Components store variations and the ability to reuse prompts using the automatically generated Variation Buttons meant that they could create, revise and evaluate a larger number of variations with ABScribe. W12 noted that storing variations in a linear flow is *"just very clunky"*, leading to a *"higher cost of keeping multiple versions"*, *"more scrolling in the document"*, and *"taking a longer time to find anything specific"*. While exploring variations with Baseline, W12 noted that: *"I sort of gave up about halfway because it was taking too long to vaguely remember, oh there was this version that I thought was good, but I wasn't able to find it because the document became so long, so I just grabbed whatever to just finish."*

F2: ABScribe enhances writers' ability to explore more granular variations in context of the surrounding passage. W1, 3, and 9 mentioned that they could work with smaller text-segments when using ABScribe more easily than Baseline. When using the chat-based LLM interface in Baseline, W3 said they had *"never even considered [editing] on a sentence level because it would be so hard to go into ChatGPT and say that in the third sentence of the third paragraph...I don't even know if it [ChatGPT] has a sense of where that is."* This sentiment was echoed by W9 who noted: *''Usually, [I] would be [editing] at least a whole paragraph and see edits from that and then paste some of those edits in [to the document] but I wouldn't put one sentence in...so there isn't very fine-grained control [in Baseline]...whereas here, [with ABScribe], because it takes less effort, I'm like okay, I can do one sentence."*

Being able to work with smaller text-segments, however, made some participants worry about the overall coherence of the passage. W9 said *"I fear that if I edit small chunks, then the tone of different chunks end up becoming different. Whereas I kind of want to change the tone of the whole thing to one specific thing."* W3 shared a similar concern but noted that the ability to view and edit variations in-place, within the context of the surrounding passage reduced their concern: *"ABScribe is vastly superior for any kind of fine-grained edits which become incredibly difficult to deal with [in Baseline] if you're trying to do different edits on smaller variations of the text, unless you want to perform a single edit and immediate make that change."*

F3: ABScribe nudges writers toward an imperative LLM prompt writing style in contrast to the conversational style in Baseline. W1, 3, 6 and 8 noted how the automatically generated Variation Buttons that captured prompts they previously wrote influenced their prompt writing style to be more direct. W1 mentioned that in ABScribe, *"you're conscious of the fact that you are designing a button, and that forces you think within that framework. You're not really asking someone to do something [like in the chat-based interface]. You're giving it instructions to*

make a button...which make things very simple, very easy, very quick...making it much easier to create prompts, and then use them subsequently, when they're being instantly converted to a new use-case." The conversational approach sometimes led users to anthropomorphize the AI, using superfluous words that do not influence the quality of the variation generated. For example, we observed that W8 would say "Can you p*lease make this variation shorter?"* when instructing the chatbot, but opt for a more direct *"make it shorter"* prompt for ABScribe. When asked why, they said the conversational AI assistant felt *"similar to Clippy"*, referencing the Office Assistant from a discontinued intelligent interface from Microsoft with an interactive animated character [19].

F4: ABScribe nudges writers toward composing generalizable and atomic LLM prompts in contrast to the complex and variation-specific prompts in the conversational approach in Baseline. W3 and 12 shared that with ABScribe, they were intentionally trying to simplify their prompt design to make them generalizable and reusable across different text segments. W3 said *"I felt more like I was going to create a generalized prompt, that I will probably reuse later. So it felt like it ought to be something that's simple that could be applied to a variety of situations rather than something that's specific to a single piece of text."* In contrast to W3, W9 found that due to the ease of prompt reuse, they were more likely to create longer prompts that would generalize to other contexts, when compared to the chat interface. *"I was okay with writing longer prompts, for example, to imitate the style of a character because it took less effort, and it was fun for me to do, and I knew I could use it again in other sentences. Whereas in [Baseline], I wasn't as excited because it would take more time to type, and I knew that if I had to reuse it, I would need to type it again."*

However, because participants were actively thinking of instructions they could reuse across different text segments, they were less likely to write prompts that were tied to specific characteristics of particular text-segments, as exemplified by W3's comment: *"I categorically preferred A [ABScribe] to B [Baseline]...the only advantage to B might be that it slightly encouraged me to have more nuance in the prompts I gave to ChatGPT...it just made me realize I could do that. I know I could do that in ABScribe too, but the button interface kind of guided me to more naturally consider simplistic prompts rather than prompts which might be more suited to particular tasks for those in conversation with a bot."* Decomposing custom prompts that only apply to a specific text segment into more general, more atomic prompts that apply across variations encouraged participants to *combine* multiple prompts by clicking multiple Variation Buttons. W3 considered mimicking the functionality of longer, more complex prompts by *"stacking"* multiple of their general, more atomic prompts to explore variations: *"I think I would also be more willing to try out a variety of permutations of different prompts, rather than trying to apply one prompt globally to an entire email, or full prompts to one or two sentences. So seeing how compound prompts might help at various points would be a lot better in A [ABScribe]."*

4.5.2 RQ2: Subjective Task Workload

Participants pointed to four aspects of the revision workload in ABScribe that led to the significant reduction in subjective task workload: reduction in clutter, ease of variation management, access to surrounding text context, and reduced context-switching during LLM use.

F5: ABScribe reduces task workload by reducing document clutter. W4, 5, 7, 10, 11, commented that the general lack of clutter in the ABScribe interface when dealing with multiple

versions affected various aspects of their writing and revision process. Specifically, the way ABScribe stores variations in-place with the ABToolbar as opposed to in-sequence in the doucument reduced clutter. W11 summarizes this in the following quote *"I think definitely the biggest difference would have been the fact that your your document is not as cluttered. Actually, it doesn't get cluttered basically because you can just switch between versions and the text is on in the same location. So that's already a huge boost in terms of not having a mess on my hand. And you notice at the very start I was already organized to think that way."* When comparing ABScribe to the linear storage of variations in Baseline, W5 and 8 noted that the linear interface inevitably led to cluttered and messy documents. W5 pointed out the difficulties of managing variations in the linear Baseline approach: *"So the linear approach was more difficult, because there was just walls of text, like they began piling up very quickly. And I tried to segment them, right, I believe, if I numbered them, it would have been better. But at the same time, it doesn't get rid of the root problem, where more and more text is, is being added to the whole draft."*

Although all participants agreed that the ABScribe interface produced a less cluttered, in contrast to what we expected, two of them perceived clutter as being not necessarily bad. W2 and W6, who were both very comfortable with the worklow afforded by Baseline as it was similar to what they were used to doing, noted that clutter didn't matter as much to them during the revision process. W2 mentioned that Baseline *"definitely"* felt more familiar, and that *"...you probably think cluttering is one of the one of the important factors to consider when people are writing, I actually don't think that's the case. That's why I don't care about whether it looks [messy during revision]."* W6, who described themselves as a *"a messy editor"*, a *"hoarder"* of various made copies of older text, felt *"really conflicted"* about the reduction in clutter because they liked being able to *"mishmash multiple versions together"* in a messy document. In describing the workflow in Baseline, they said it was *"more familiar to me than like doing it the way that you would in ABScribe. Even though in an abstract way, the non-linear approach makes a lot of sense...I feel like it just feels like that's how the design should be...but I feel like [I'm] a messy editor. And so it's, it's almost easier for me to edit, in Baseline."*

F6: ABScribe reduces task workload by enhancing variation management. Two major recurring activities in the exploration of multiple variations were variation storage, or tracking variation history, as W7 referred to it, and the comparison of multiple variations. W3, 6, 7, and 11 noted that the non-linear storage of the ABScribe interface necessitated less overhead to manage and revise multiple variations. W3 commented that by having the ABScribe interface manage storing variations for them, they were more able to focus on the writing task: *"You're fully focused on the writing changes you're trying to make, as opposed to managing the state and managing your document and managing like, that kind of stuff. So that was the biggest difference for me. So I really liked that feature. And that made a huge difference in general to the task. But because I'm less focused on management of things, or management of my thoughts a little bit, it's a lot easier just using that, like the versioning system."*

While W3 touched on the ease of variation storage, W11 discussed how the Hover Buttons enhanced the ease with which they compared variations: *"the feature made it easier to to do the comparisons, because then you can click the version that you're - you're comparing with, and then hover and look at the text [for the other variations]. Whereas, with the Baseline, you have to both keep track of where the version you're comparing with is and also simultaneously figure out which*

version you're comparing to...so that is trickier than dealing with the new approach."

Some users noted that there were some instances where comparing larger text-segments was easier to do in-sequence, and suggested that the Variation Accordion interface could also serve the purpose of in-sequence comparison. W3 notes this: *"Yeah, I mean, one nice thing about Baseline is that I do get to see all of the versions together. So they're all listed for me...but it's much harder to have a number of sentences, which you're generating different versions for because even once you start to hit two different sentences, [and] we're trying to generate different variations, the document becomes very cluttered, and becomes difficult to manage. And you forget what the context is for each of those different variations. So for - for context, and for clutter, I think A is vastly superior."*

F7: ABScribe reduces task workload by showing variations in context of the surrounding text during manual editing and LLM use. The need for considering context during the revision process came up during several interviews. W3-8, and 10 commented that thanks to the in-place comparison of variations, they were able to see what a variation looks like in a paragraphs, as exemplified by W10's comment: *"I like the nonlinear version, because when you hover on the different buttons, you can directly see the impact of different variations within the paragraph or within the context. So in that way, you know, whether the text fit into the original document or not. Whereas in Baseline, if you put all the variations linearly in the document, at some point, you just start to lose a sense of what's the context of this of this sentence, what am I writing there? Also, the chatbot in Baseline, I'd say it's pretty much [the] same as ChatGPT. So if I want something, I need to scroll back to try to look for it. So that's pretty much similar to the current AI writing system."*

W7 echoes this preference for the non-linear ABScribe interface: *"When you're writing a paragraph, you're not looking at a sentence in isolation. So if you're changing a particular sentence, you want to see how it looks in comparison to the rest of your text. And so to have the nonlinear version allows you to kind of do that more seamlessly than with a linear version, where you'd have to reorganize a lot more in order to have that effect."* Whereas, W10 mirrored this, and noted that manually organizing variations while simultaneously figuring out context was challenging: *"I need to manually think of a way to organize all the variations so that I understand what they mean. Or like, what, how they're connected to the original text. That cost a lot of time. And it's like, very high physical demand."*

F8: ABScribe's variation storage and in-place LLM revision reduces subjective task workload . Interactions that brings the user outside of the primary text editing interface and break their flow of writing or revising were perceived to be effortful and time-consuming. W7 points out how the baseline interface leads to these sorts of interactions: *"So if I were to go into version history, and I wanted to go back to a very particular change in one particular paragraph, but I made that like 50 changes ago, I would either have to revert back to something where all of the document would have been unchanged, or I'd have to do like a very inconvenient and kind of cumbersome process of like copying that particular change from that particular version history into my current doc and then proceeding, which is, like tedious"* W4, 5, 7, and 9 commented that they all had various ways of managing different variations of text when using the baseline or outside of the study that required them to leave the primary text editing interface to perform comparison of variations. This was either to copy paste different versions from separate documents into the primary text editing interface for comparison in sequence, or simply to compare variations side by side in separate documents. W5, 7, and 9 all found the ABScribe interface less cumbersome, especially when performing edits on several

smaller text segments, due to the lack of context-switching. W9 explains this here: *"Whereas here [ABScribe], I think because it takes less effort, like okay, I can do one sentence. I also want to do another one. So I'll do that. I don't need to copy the whole paragraph in [the chat-interface of Baseline] and try to get an answer from that. I can just do it to those sentences."*

F9: ABScribe reduces task workload by making prompts more reusable than Baseline. Almost all users (W2, 3, 6-10, 12) noted that prompt reuse was much easier in the ABScribe interface. W3 commented on how the baseline interface imposed a *"memory load or cognitive load issue to remember what prompts you have."* W2 notes how the AI buttons of the ABScribe interface eliminate the need to rewrite prompts stating *"I don't need to rewrite prompts every time. It was really very quick and efficient. The usability of this one in terms of buttons, the reusing the prompts is very good."* While W8 called ABScribe's AI buttons *"much more streamlined."* W3 sums up the interaction concisely in the following quote: *"A [ABScribe] is vastly superior for reusability, there is no question. B [Baseline], you basically have to work from memory, which can also be fine. But with version A, you click it, you don't have the same memory load or cognitive load issue to remember what prompts you have before, cannot be more better facilitated."*

4.6 Discussion

In this work, we present the design and implementation of ABScribe which is composed of five key design elements: Variation Components, Hover Buttons, the Variation Accordion, AI Buttons, and AI Insert. Our comparative evaluation study shows that these elements reduce task workload and significantly enhance user perceptions of the revision process when managing multiple variations of text segments in human-AI co-writing tasks as compared to a familiar baseline editing system (Section 4.4).

We present six findings (F1-6, Sections 4.5.1 and 4.5.2) that provide evidence for the efficacy of our design elements. These findings offer insights into how writers utilize both chat-based and in-place Human-AI co-writing interfaces (Sections 4.2.3, 4.5.1). They also illustrate the affordances and limitations of these interface types for exploring multiple writing variations, their influence on the size, granularity, quantity, and diversity of variations, as well as the prompt style users take on.

4.6.1 Non-Linear Text Revision Control

In ABScribe, the Variation Component, and dynamically placed Hover Buttons provide an effective approach to managing text variations non-linearly. This offers an alternative to the linear text revision control features found in current editing interfaces and chat-based AI-assistants.

We find distinct influences on task workload (Section 4.5.2) and the style of prompts created by users between these two methods (F3-4, Section 4.5.1). Our interface is grounded in a non-linear nature of the writing and editing process [325, 301], offering a way for designers to support multiple fine-grained variations without overwhelming users (F2, Section 4.5.1). As our design for the Hover Buttons and Variation Components builds upon a familiar rich-text editing interface, they can potentially be integrated into existing document editors without major layout changes. This would enable writers to work more closely with LLM-based generative AI content, within the context of the surrounding text (F7, Section 4.5.2).

We also observed that the non linear approach affords advantages such as viewing, editing, and combining multiple variations together. Notably, participants W1 and W4 highlighted the usefulness of the non-linear text revision control even without AI (F6, Section 4.5.2), indicating the broader relevance of the Hover Button and Variation Component interface overlay for the wider range of text editors.

4.6.2 Scaffolding Prompts Focused Around Specific Writing Tasks

Our design for AI Buttons demonstrates how scaffolding LLM prompts around interface elements can influence user prompt-writing behavior. Specifically, the AI buttons, auto-generated after writers create a prompt for a text-segment variation, encourages writers to craft more direct, imperative prompts (F3, Section 4.5.1). It also made them reflect more on the revision process, shifting the focus from conversing with a chatbot to designing a button that represents their writing style for reuse across variations (F9, Section 4.5.2). We noticed a trade-off: while making prompts more reusable, it nudged users away from conversational prompts tied to a specific variation's nuances (F4, Section 4.5.1). For example, they were more polite in the chat-based interface and direct when using the button scaffold. This highlights the importance of considering how different UI scaffolds influence the kinds of prompts users write.

These findings contribute to the body of research on encouraging writers' experimentation of LLM output using modular LLM interfaces [117], and expands the scope of valuable design paradigms in text-editing, such as the use of reified text selections [116, 115, 20], to supporting the exploration of multiple variations in LLM-based revision-focused writing.

4.6.3 Moving Beyond Systematic Exploration to Systematic Evaluation of Variations

Our interface elements could be further developed to support the systematic evaluation of variations. While we want to avoid premature elimination of rough ideas [112, 42], we also want to afford critical assessment of variations in relation to each other [345, 42]. For instance, if we develop an extension that lets writers test different versions of the ABScribe draft containing a randomized subset of variations, they could link our interface to an open-source A/B testing framework like MOOClets [284], UpGrade [233] or Planout [16]. This would allow writers to quantitatively evaluate which versions best meet their objectives based on specific metrics. A copywriter, for example, wanting to rapidly explore and assess different advertisement versions, can use ABScribe to explore draft variations, compare variations in the draft's context using Hover Buttons, select a subset for further evaluation, and run randomized online experiments via an A/B testing framework. Moving beyond systematic exploration, which our current study covers, towards systematic evaluation is a logical next step. This opens a diverse design space for A/B authoring tools that simplify designing variations for evaluation, making it a promising avenue for future research.

Our design seeks to enhance the natural flow of writing and editing texts by supporting close collaboration between humans and AI, grounded in established theories on the revision process. As highlighted in Section 4.2.1, revising text is not merely about superficial changes [91], it delves into deeper subprocesses like idea formulation [92] and meaning discovery [325]. Skilled writers view revision as a cyclical and recursive process [325, 301], diving into the content repeatedly with

different aims such as shaping a persuasive narrative [325], playing with word choices and tone [137], and switching between various writing tasks during text revision [87]. Our interface provides a concrete example of affordances that editing interface designers can leverage to make the Human-AI collaboration in revisions more congruent with established revision theory.

In Section 4.2.1, we hypothesized that improving ease of use for handling multiple variations could help us effectively apply HCI design principles, affording the user greater freedom in experimenting with variations, thereby avoiding premature fixation on an initial idea [345, 151, 345]. As we found in F1 (4.5.1), the design of ABScribe helps writers to not commit to an initial idea too early, and instead, explore a greater quantity of variations before evaluating them, providing evidence for the applicability of parallel exploration in text revision.

4.6.4 Limitations

Our work has two limitations to external validity, which are common in lab-based evaluation studies. First, our evaluation study was restricted to English writers. Although the underlying LLM, GPT-4, supports multiple languages [249], suggesting that our interface can extend beyond English, our study focused on English writing tasks. Without specifically studying how different languages influence task workload and user perceptions of the revision process, we cannot comment on the wider applicability of our tools, and would caution designers against directly applying our design without further evaluation. Second, our evaluation study was limited to a single 1.5-hour session with two guided writing tasks to ensure we could conduct our comparison in a controlled setting. We crafted realistic writing tasks that could be completed within the study's timeframe, representing use cases from prior studies and commercial AI apps. Ideally, writers would select their own writing task and spend an extended period, possibly spanning several days or even a week, to revise and explore variations.

To address some of these concerns, we complemented the writing tasks with in-depth user interviews, allowing writers to reflect on and discuss the implications of our design beyond the study's writing tasks. Several participants, such as W1, 3, 4, 11, and 12, expressed interest in using our tool for creative and academic writing, in settings beyond the use cases we explored.

4.7 Conclusion

In this work, we presented ABScribe, a human-AI co-writing interface built for swiftly exploring multiple writing variations using Large Language Models (LLMs).

Our interface is composed of an ensemble of five distinct elements: Variation Components, Variation Accordion, Hover Buttons, AI Buttons, and AI Insert. Collectively, these elements not only markedly decrease task workload ($d = 1.20, p < 0.001$) but also bolster user perceptions of the revision process ($d = 2.41, p < 0.001$), in comparison to a popular AI-integrated editing workflow consisting of a rich text editor augmented with a chat-based AI assistant and the ability to insert AI generated content.

Our evaluation with writers (N=12) validate the efficacy of our design and offer insight into how writers leverage LLMs to explore variations, revealing a preference for non-linear over linear revision strategies, especially when engaging with a multitude of variations at finer granularity levels. We also found that scaffolding LLM use with task-focused UI components, like buttons, encouraged

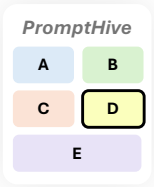
writers to create more generalizable prompts and use more direct, imperative language in prompt design. Our work informs HCI research on the design of Human-AI writing interfaces for the rapid exploration of writing variations.

ID	**Writing Experience**	**AI-Assisted Tools Usage**
W1	***Moderate:*** Worked as a staff writer for two political science publications. Writes fiction as a hobby.	***Moderate:*** Uses ChatGPT to edit writing projects, as well as receive feedback and suggestions for further passages, primarily after completing a passage to identify areas for further revision.
W2	***Moderate:*** Taught ESL courses to non-native English speakers, specializing in IELTS, TOEFL, and business English instruction.	***Experienced:*** Uses ChatGPT by feeding it the main points of the article to generate a draft, and then editing the responses provided by ChatGPT.
W3	***Highly Experienced:*** Has experience writing papers, specifically about writing tools for HCI. Has also, published a novel, and has publications in many highly regarded literary venues (BOMB, LitHub, FENCE, and more). Participant also writes their own music.	***Moderate:*** Uses ChatGPT for drafting messages, seeking feedback on fiction, and drafting small sections of research papers. Has experience with Respondable, a service in the Gmail plugin called Boomerang, for writing emails using AI.
W4	***Highly Experienced:*** Writes fiction and published one novel, some sci-fi and fantasy short stories, and several articles for blogs, magazines, and satirical news sites. Worked as a staff writer as an undergraduate, a professional screenwriter for two independent studios. Also teaches two first-year writing classes in a liberal arts college. Achieved a MFA in Creative Writing.	***Limited:*** Briefly experimented with ChatGPT to test its capabilities by asking it to write some scripts, essays, and articles. Found the results to be amusing, but lacking in perspective and personality.
W5	***Experienced:*** Writes content for social media profiles for an NGO. Studied English Literature during both bachelor's and master's degrees. Writes music, having penned 65 songs, and promotes it through social media and music platforms such as YouTube and Spotify.	***Experienced:*** Used Grammarly for on-the-go editing to write and ChatGPT much more extensively for both idea generation, as well as summation and synthesis of large bodies of text. Also found ChatGPT useful for helping figure out parts of creative works that may feel like they have gaps which can be prone to miss.
W6	***Experienced:*** Worked for two national English language newspapers, including contributions to their weekend magazines, kid's sections, international section, in addition to also publishing fictional short stories for the newspaper. Completed a Creative Writing Certificate, and currently primarily writes about research.	***Limited:*** Used ChaptGPT in a very limited capacity, mostly to brainstorm assignment structures and topic sentences when writing.
W7	***Experienced:*** Writes academic articles, general interest articles and reviews for local newspapers. Also writes short stories for sharing with friends, and has a short story published in an locally published anthology. Expertise is primarily in creative nonfiction.	***Limited***: Briefly experimented with ChatGPT.
W8	***Moderate:*** Written mostly technical papers, but also wrote some short stories as a hobby.	***Moderate:*** Uses ChatGPT every other day mostly to proofread, create templates for texts, and find the right creative direction when writing.
W9	***Experienced:*** Usually writes research papers in computing education and blogs. Blogs usually cover personal experiences at work as well as hobbies.	***Limited:*** Used Grammarly and ChatGPT to edit writing.
W10	***Moderate:*** Wrote some column articles for personal social media accounts, and several research papers over the past five years.	***Experienced:*** Uses Notion.AI, ChatGPT, and GPT-4. Uses Notion.AI for generating bullet points and brainstorming ideas, ChatGPT for generating templates for writing, and sometimes summarizing related work for research purposes.
W11	***Highly Experienced:*** Focuses on academic writing such as papers, scholarship applications, reviewing, etc.	***Advanced:*** Has experience with Grammarly, Notion, Obsidian with GPT plugins, and ChatGPT. Mostly uses these tools to clean up sentences, and sometimes uses them to brainstorm titles for papers.
W12	***Experienced:*** Engages in hobby novel writing, academic writing, blog writing	***None:*** Has not used AI-assisted writing tools, but has experience with AI image generation using written prompts.

Table 4.1: Self-reported experience with writing and using AI-assisted tools. Expertise labels for writing range from very limited to highly experienced, and labels for the AI tools ranges from none to advanced. Details on prior writing experience and AI tool usage is also included for each participant.

Chapter 5

The PromptHive System



> Smashing the red button on loop saying yep, yep, yep — that doesn't seem like human control to me.
>
> *Shashank Joshi [122]*

Research Context: The previous chapter examined how LLMs can be used to accelerate content variation exploration through interface elements for parallel editing. However, as AI becomes more integrated into educational content creation, critical questions arise regarding how to keep human subject matter experts at the forefront. That is the focus of this chapter, addressing our fourth and final research question:

> *Q4: How might we retain domain experts' sense of agency and control in AI-assisted educational content authoring as they explore multiple alternative instructions to guide AI output?*

The answer is to engage domain experts not as passive evaluators who merely accept or reject AI output but as active decision-makers who experiment with prompt variations to refine AI-generated content within their instructional context. We demonstrate this through the design of PromptHive, extending our investigation in ABScribe's single-author AI collaboration to a multi-expert setting. This work was a collaboration with the CAHL group at UC Berkeley.

Related Publication:

Mohi Reza, Ioannis Anastasopoulos, Shreya Bhandari, and Zachary A. Pardos. 2025. *PromptHive: Bringing Subject Matter Experts Back to the Forefront with Collaborative Prompt Engineering for Educational Content Creation.* In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 148, 1–22. https://doi.org/10.1145/3706598.3714051

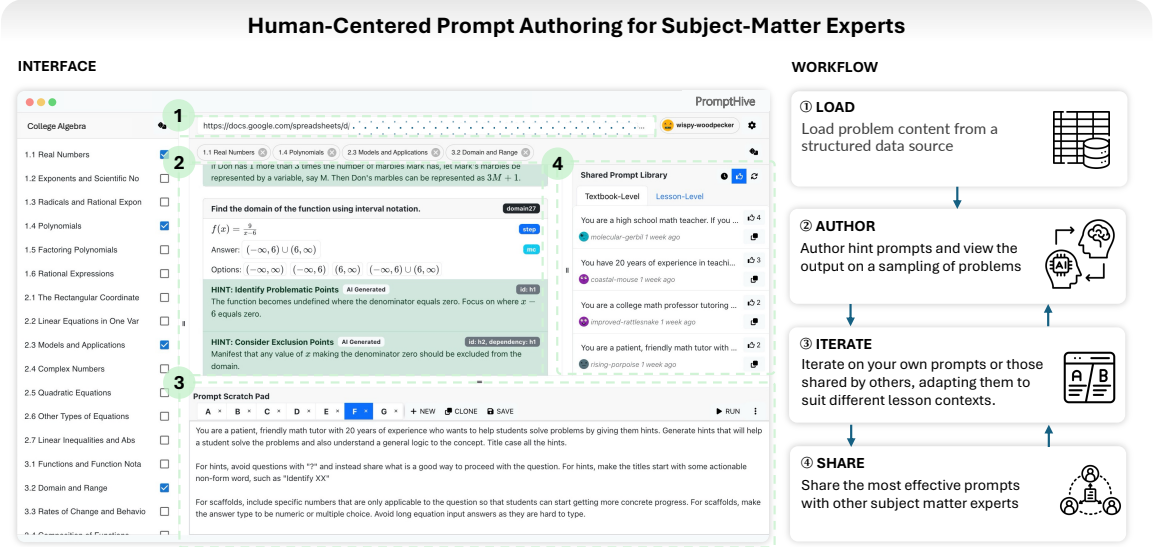


Figure 5.1: The **PromptHive** Interface and Workflow: (1) **Load:** Import textbook lessons and problems by pasting a link to a structured data source. (2) **Author:** Create hint prompts and view the output generated for a variety of problems from different lessons using sampling buttons. (3) **Iterate:** Refine your own prompts or those shared by others by cloning them into the scratchpad, experimenting with different changes, and evaluating the impact by comparing output variations using buttons labeled A, B, C, etc. (4) **Share:** Save effective prompts in a shared library for other subject matter experts to clone, evaluate, and modify for different lesson contexts.

Abstract: Involving subject matter experts in prompt engineering can guide LLM outputs toward more helpful, accurate, and tailored content that meets the diverse needs of different domains. However, iterating towards effective prompts can be challenging without adequate interface support for systematic experimentation within specific task contexts. In this work, we introduce PromptHive, a collaborative interface for prompt authoring, designed to better connect domain knowledge with prompt engineering through features that encourage rapid iteration on prompt variations. We conducted an evaluation study with ten subject matter experts in math and validated our design through two collaborative prompt writing sessions and a learning gain study with 358 learners. Our results elucidate the prompt iteration process and validate the tool's usability, enabling non-AI experts to craft prompts that generate content comparable to human-authored materials while reducing perceived cognitive load by half and shortening the authoring process from several months to just a few hours.

5.1 Introduction

As Large Language Models (LLMs) push the boundaries of what computers can help humans create, the question of how to design authoring interfaces that effectively connect domain experts with the prompt engineering process becomes increasingly salient. With the right design [346], such interfaces could enable experts to steer the output of LLMs toward content that better aligns with the nuances and needs of their domains and transform the role of the subject matter expert from a *producer* to a *curator* — a competent and critical judge who *instructs* the AI agent on what is needed, *evaluates* the output, and *iterates* on the instructions until the results are satisfactory. Instead of replacing human experts, these interfaces could help bridge human intelligence with machine intelligence to dramatically reduce the time and effort required to create content that adheres to expert standards.

To realize the *producer-to-curator* shift and integrate domain expertise more closely into prompt engineering, we need authoring interfaces that: (i) deeply embed LLMs within existing expert workflows, augmenting content creation with carefully scaffolded interface support for prompt engineering; (ii) encourage experimentation on many prompt variations to systematically test the impact of changes in instructional wording on model output; (iii) offer mechanisms for curating prompt formulations that work well at various levels of abstraction; (iv) integrate generation into the publishing workflow. However, designing authoring interfaces that support experts across all four fronts is difficult as LLMs pose unique usability challenges tied to high metacognitive demands during prompt construction [338], and users can struggle to get the models to integrate well with their existing workflow as even small perturbations such as adding a space at the end of a prompt can cause the LLM to change its output [293]. For domain experts who aren't AI specialists, recent literature on prompt engineering has also highlighted how designing effective prompts can be surprisingly difficult [381, 31].

In this work, we explore ways to enable domain expert-driven prompt engineering and answer empirical questions on how they permute and evolve prompts through the design of **PromptHive** (Figure 5.1), an open-source collaborative prompt authoring interface that we apply to the content authoring workflow of an open-source adaptive tutoring system, OATutor [261], to support domain experts with writing prompts for generating hints to homework problems. PromptHive encourages rapid experimentation and systematic testing of alternative prompt variations via a set of randomization features for sampling problems from a structured data-source, as well as buttons that pair prompt variations in the scratchpad with corresponding model output for easier comparison. This approach has been shown to be effective in a recent Human-AI authoring interface in a different context [282]. To support prompt curation, PromptHive features a shared library where users can save, clone, and upvote useful prompt formulations at two different levels of abstraction - the textbook-level and lesson-level. PromptHive also features a back-end logging engine that collects rich user interaction data on how prompts evolve as users create, modify, and share them.

To validate our design, we conducted two studies: (i) a three-stage user study with ten subject matter experts who had prior experience in manually authoring hints using an expert workflow augmented by PromptHive; (ii) a learning gain study with 358 learners, comparing the efficacy of hints generated using PromptHive with hints previously authored manually by subject matter experts. Our findings provide rich empirical data on how users collaboratively iterate on prompts to generate hints for an entire college-level algebra textbook and show that users can successfully use PromptHive to create hints that are on par with those authored manually by subject matter

experts who did not make use of generative AI, while cutting mental workload by *more than half* and the time required from several months to just a few hours. The subject matter experts rated PromptHive as having excellent usability (SUS score = 89/100) and felt that it fostered trust in the performance of the LLM model (Figure 5.5). In this work, we contribute:

1. The PromptHive interface and workflow that integrates subject matter experts into the prompt-engineering process through collaborative iteration on prompt variations, combining human expertise with the efficiency of generative AI. Our implementation works with a tutoring system to generate hints for problems from various open educational resources and is *fully open-sourced* to enable extensions to other systems and domains.

2. Empirical results from two studies that validate our design: a user study with ten mathematics subject matter experts and a learning gain study with 358 college students. These studies demonstrate the efficacy of PromptHive and its advantages over manual content creation, reducing authoring time by a *factor of 30 while maintaining expert-level quality.*

3. A logging engine that *captures rich user interaction data* on how subject matter experts permute and evolve prompts while using PromptHive. This engine can serve as a valuable resource for future HCI research aimed at gaining deeper insights into the collaborative prompt-refinement process.

5.2 Related Work

We review the literature in HCI and generative AI to explicate the need for integrating subject matter expertise into prompt engineering from both ethical and practical standpoints. We also highlight the challenges associated with designing human-centered interface support for prompt engineering and examine existing systems for collaborative prompt engineering to situate and distinguish PromptHive from other tools.

5.2.1 Integrating Subject Matter Expertise into Prompt Engineering

Placing subject matter experts in the driver's seat of prompt engineering is crucial as they possess the judgement necessary to evaluate the output of LLMs in their domain [294, 186]. Providing subject matter experts with tools to work with data related to their domain can also empower them and assist them in further iterating upon their workflow [153]. Furthermore, software engineering teams may not be representative of the users and their demographics [1], and as a result, may be less capable of producing prompts that address the diverse needs and perspectives to which experts are more likely to be attuned. Domain expert-driven prompt engineering is also important because of the wide-ranging impact LLMs are having across many, many domains including art [173, 386], medicine [384, 224], law [120, 88], and education [184, 258]. Unlocking the potential of pretrained generative models hinges on aligning them with human intentions [360]. Researchers, organizations, policy makers, and society at large will need to grapple with the question of how to best involve human experts in AI-intensive workflows. Preliminary studies have shown the potential benefits of expert involvement in the application of LLMs. For example, Kumar et al. found that an instructor-tuned LLM significantly boosted student interactions with a chatbot compared to plain ChatGPT as a

baseline [184], and Wang et al. extended the Mixture-of-Experts Paradigm to prompt optimization, demonstrating how breaking up a problem space into sub-regions controlled by specialized human experts can help with prompt optimization [360].

In this work, we contribute an open-source system that exemplifies how to effectively integrate multiple subject matter expertise in the context of educational content creation and contribute to the ongoing discourse within the HCI and Human-Centered AI communities on retaining human control while increasing automation [314]. In contrast to Sheridan and Verplank's characterization of automation and human control as a unidimensional spectrum [309, 316], we adopt Shneiderman's two-dimensional HCAI framework [316] and explore ways to increase automation *without* encroaching upon subject matter experts' sense of control and trust over the content.

5.2.2 Usability Challenges of Designing Prompt Engineering Interfaces

Tankelevitch et al. notes that the same unique properties that make LLMs powerful, such as their flexibility across multiple input/output spaces and generality across tasks, pose usability challenges for the design of human-centered Generative AI systems due to the high metacognitive demands of prompting [338]. These usability challenges indicate that subject matter experts need interface support to carefully scaffold the task of prompt writing, allowing them to focus on the core value they bring–the ability to guide the LLM on what is needed and evaluate output quality. Beyond addressing usability challenges, such scaffolding must also account for the inherent unpredictability of generative models [293] and make experimentation and iteration a core part of the prompt design process [282, 164].

In PromptHive, we design and validate explicit interface support for iterating on multiple prompt variations in the form of prompt-output buttons that pair prompts with the generated output in a scratchpad interface to allow experts to qualitatively compare the impact of changes to model output. We leverage the latest advancements in LLMs to deal with model unpredictability, such as the ability to reliably adhere to specific output schemas in GPT-4o [250], self-consistency prompting [361], and multimodal capabilities that take into account images in problem content, to effectively integrate PromptHive within a real-world expert workflow for generating hints in a production tutoring system [259].

5.2.3 Designing Effective Content Authoring Tools for Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have historically required highly time consuming authoring processes. Early iterations of authoring tools required 200-300 hours of use to produce a single hour of instructional content and required experience with programmatic interfaces [4]. Development of the Cognitive Tutor Authoring Tools (CTAT) quickly reduced this time to 50-100 hours to produce one hour of instructional content [365]. CTAT provided content authors with a GUI which allowed for content authoring with minimal knowledge of programming and coding skills. The GUI allowed for content creation to take place with the respective interface, which greatly supported independent content creation processes [4].

Following the example of CTAT, many future ITS and ITS-like systems supported their content creation ecosystems with sets of builder tools that were focused on providing content authors an easy-

to-use interface, usually in the form of a GUI. One example of such tools is that of the ASSISTments builder. While not fully an ITS, ASSISTments is an adaptive tutor that places the instructor in an important role within its system and simplifies the content authoring process [129, 279]. To accomplish this, the builder provided easy integration of problem-help features such as hints and scaffolds, while also allowing skill mapping of knowledge components. Furthermore, it managed to match the lower end of CTAT's 50 hour estimate of development for one hour's worth of instructional content despite using a GUI instead of a programmatic interface [279]. The ASSISTments Builder also supported problem variabilization directly through its GUI, allowing for easier variation of problem content [279].

While GUIs have provided many benefits for content authoring, there are still many challenges that accompany them. With the large number of tutoring systems available, it can be difficult for teachers to have to learn a new content authoring environment every time they wish to utilize a new tutor [365]. Furthermore, creating problems with more complicated structures may not just be difficult, but in some GUIs may not even be possible [279]. This has resulted in difficulties for new systems to balance between efficient and easy content creation while also supporting more complex content creation. When compared to the GUI of the ASSISTments builder, the programmatic spreadsheet-based authoring interface used in OATutor did not result in any significant differences regarding time taken for content authoring, but showcased higher accuracy when curating content, albeit with a considerably lower usability score [307].

These challenges underscore the need for the next generation of LLM-infused content authoring systems to integrate seamlessly with *existing* expert workflows, ensuring that AI adapts to the needs of users, rather than the reverse. In this work, we demonstrate how systems like PromptHive can successfully augment complex workflows involving multiple experts with minimal disruption.

5.2.4 Generative AI in Tutoring Systems

In recent years, LLM improvements have had a significant influence on tutoring systems and educational contexts. Evaluations of ChatGPT's decimal skills indicated that it can respond accurately to conceptual questions but struggles more with respect to number lines and decimal point problems [239]. Furthermore, when examining student answers, ChatGPT accurately assessed the correctness of seventy-five percent of them, while generating feedback that was similar to that of instructors. ChatGPT's ability to produce informative worked solutions proved effective for learning in Algebra and Statistics subjects [258]. In higher education, across 53 studies comprising 114 question sets and over 49,000 multiple-choice questions (MCQs), ChatGPT 4.0 correctly answered 75.5% of all MCQs, achieving a passing score on the majority of the problem sets [237]. It has also been shown to effectively evaluate the quality of learnersourced MCQ distractors [228].

LLMs have seen usage with respect to tutor question quality evaluation. Generative Students is a prompt architecture that utilizes LLMs to simulate student profiles for the purpose of simulating believable MCQ answers [211]. Generative Students demonstrated a high correlation between how real students responded to the sample question and how the simulated students did. Furthermore, the system demonstrated that an instructor could improve their question quality using these simulated students. A similar approach was also used to learn question difficulty parameters using LLM-Respondents, thereby allowing for appropriate quality tutoring questions to be selected [209]. Beyond simulated students, LLM tutoring use also has taken the form of chatbots. An ASSISTments

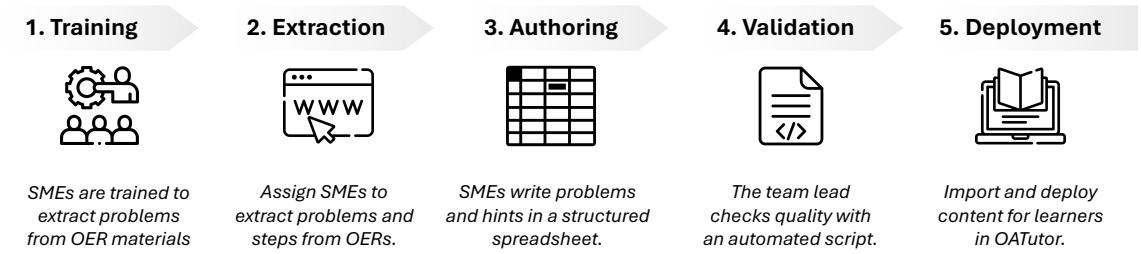


Figure 5.2: A structured workflow followed by Subject Matter Experts (SMEs) for creating manually authored content in OATutor. The process includes: (1) Training SMEs to extract relevant problems from Open Educational Resources (OERs); (2) Assigning SMEs to identify problems and steps in OERs; (3) Authoring structured problems and hints in a spreadsheet format; (4) Validating content quality using automated scripts led by the team lead; and (5) Deploying the finalized content into OATutor for learner engagement.

integrated chatbot utilizing GPT 4.0, while not providing statistically significant learning gains, increased the positivity of students' attitudes, even though they displayed a lower confidence of solving a similar problem after the chatbot help intervention [51].

LLM-generated questions themselves have also been evaluated. When compared against questions from a published Creative Commons textbook, there were no statistically significant differences between the difficulty of algebra textbook questions and similar ones generated by ChatGPT [23]. These results indicate that ChatGPT is capable of producing algebra problems of similar quality to those from textbooks.

While much work has been done on tailoring the output of these tools for learners, these studies do not interact with the role of content authors and subject matter experts within these forms of content generation. The above studies focus mostly on content generation through prompts, bypassing discussions of subject matter expert involvement (and thus potential improvement in content quality). These and other studies on similar generative AI tools have overlooked this dimension of user interaction situated with respect to giving user's agency to co-design with generative AI's internal mechanics [273]. None of the aforementioned papers provided subject matter experts to engage in the prompt creation process, accentuating the gap between subject matter experts and the content creation process.

5.3 Designing PromptHive

In this section, we describe the expert workflow that PromptHive is designed to support, the core design requirements, the development process, and key interface elements.

5.3.1 Understanding the Expert Workflow

To understand how generative AI could support educational content creation, we explore the details of the existing expert workflow within the OATutor project. Open Adaptive Tutor (OATutor) was selected for its Creative Commons content library, rapid experimentation capabilities, structured json-based content format, and documentation of the expert-authoring of its content [7]. Tutoring systems based on ITS principles have proven effective for learning [256], as has OATutor in the

subjects of Algebra and Statistics [258]. The lead author conducted a series of in-depth conversations with the head of OATutor's content team, who oversees the training of subject matter experts (SMEs) in content authoring. As outlined in Figure 5.2, the expert workflow that has previously been employed in OATutor followed a five-step process:

1. **Training:** SMEs with prior tutoring experience in different subject areas, including mathematics, physics, or chemistry were trained by the content team lead on how to extract problems from open educational resources (OERs) such as free textbooks from the OpenStax project [327]. An introductory Canvas course was provided to the SMEs to prepare them accordingly.

2. **Extraction:** Then, SMEs were assigned specific chapters or topics from the OER materials and tasked with curating problems, breaking them into smaller, manageable steps.

3. **Authoring:** The problems were entered into a structured spreadsheet. Each row represented a problem and its associated steps, and SMEs authored two types of hints for each step: plain hints without answers and scaffolded hints that provided step-by-step solutions. The spreadsheet required adherence to strict formatting rules, including custom math formula formats, and this entire authoring process was done manually.

4. **Validation:** Senior content team members conducted quality checks using an automated validation script to ensure adherence to the correct structure and format in addition to manual checks for correctness and aspects such as spelling and grammar.

5. **Deployment:** Once validated, the content was imported into OATutor, making it available for students. Once on the system, feedback could be received regarding the problems, which could lead to further iterations and fixes.

5.3.2 Eliciting Design Requirements

To effectively integrate generative AI into this expert workflow, we surveyed frameworks for Human-AI collaboration and prompt-authoring interfaces for LLMs to derive an overarching design requirement (R0) and five supporting design requirements (R1-5) for PromptHive:

- **Control (R0):** Ensure SMEs retain control while leveraging automation. Drawing on the two-dimensional Human-Centered AI (HCAI) framework by Shneiderman [316], we aimed to automate the content generation process as much as possible without diminishing SMEs' sense of control. To achieve this, we provided interface tools that allowed SMEs to easily guide content generation and oversee quality, while automation reduced the effort required to produce educational materials.

- **Integration (R1):** Seamlessly integrate AI support within the existing expert workflow. The system needs to read and write content in the same format used by SMEs and preserve human oversight over the AI-generated content. Integration with current processes is crucial to ensure smooth adoption without disrupting established workflows.

- **Simplicity (R2):** Given the usability challenges associated with human-centered generative AI, particularly the high metacognitive demands of prompt engineering [338], a key requirement is to minimize cognitive load by scaffolding the prompt-writing process. The system should

enable SMEs to focus on content quality — such as instructional clarity and hint structure — rather than dealing with technical complexities like API management or formatting rules.

- **Trust (R3):** Building trust in AI is a central challenge identified in the explainable AI (XAI) literature [10]. Therefore, a core requirement is to foster trust in the system by providing SMEs with tools to evaluate and validate AI-generated content thoroughly, ensuring alignment with their educational goals and quality standards.
- **Iteration (R4):** Given the inherent unpredictability of LLM outputs, experimentation has been recognized as a critical requirement for effective prompt-authoring [282, 164]. PromptHive must support rapid iteration on multiple prompt variations, enabling SMEs to compare outputs and refine prompts efficiently to achieve optimal results.
- **Collaboration (R5):** Finally, given that the existing expert workflow involved multiple content team members, in alignment with the social paradigm of prompt design [363], we seek to support collaborative prompt engineering to write and refine effective formulations together.

5.3.3 Brainstorming & Design Review Sessions

To transform these design requirements into concrete interface support for subject matter experts in PromptHive, the lead author held weekly brainstorming and design review sessions with the content team lead over a span of six months and engaged in rapid prototyping, transitioning from low-fidelity sketches to the fully-functional open-source system that this paper contributes. The design process involved three key stages: (i) Paper-Prototyping: Low-fidelity sketches explored ways for SMEs to integrate human-authored content with generative AI and rapidly iterate on hints; (ii) Cognitive Walkthroughs: Informal sessions were held to refine the paper prototypes and ensure the designs were intuitive; (iii) Web-Based Prototyping: High-fidelity prototypes were built based on the refined designs. There was significant overlap and back-and-forth iteration throughout these stages, with feedback from walkthroughs informing subsequent changes.

5.3.4 Developing the PromptHive Interface

This iterative process led to the development of an interface that supports a 4-stage prompt-authoring workflow at two levels of abstraction, as described in Figure 5.3:

1. **Load:** To fulfill R1 and enable seamless integration between human-authored and AI-generated content, we implemented a loading mechanism in PromptHive that allows SMEs to import content from structured data sources, such as an OATutor spreadsheet, by pasting a link.
2. **Author:** Users can author hint prompts and view its output on a sampling of problems. To address R2 and R3, and ensure SMEs can thoroughly test prompts against a variety of problems and lessons even when the content pool is large, we introduced randomization buttons. These allow users to systematically sample problems for more comprehensive testing.
3. **Iterate:** To fulfill R4 and support rapid iteration on prompt variations, the "Prompt Scratch Pad" interface enables SMEs to edit and experiment with multiple prompts in parallel. Users can clone prompts to test variations, compare generated outputs side by side, and quickly

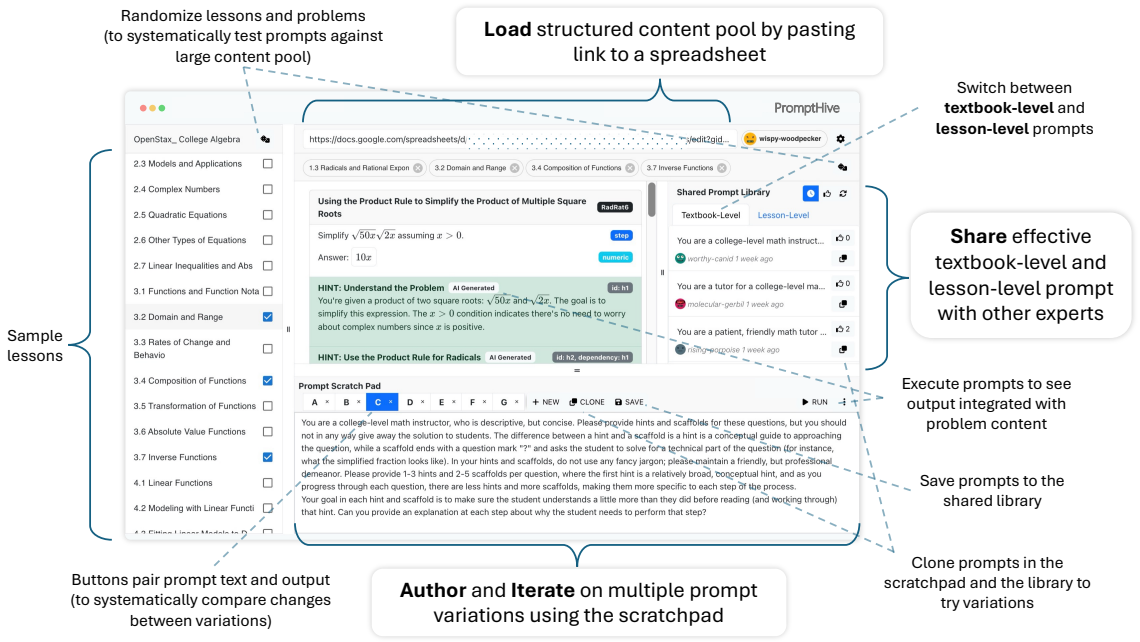


Figure 5.3: An overview of how the interface elements in PromptHive align with the 4-stage, 2-level workflow for educational content development. The workflow includes loading a structured content pool (via a spreadsheet link), authoring and iterating on prompt variations using the scratchpad, randomizing lessons and problems for systematic testing, and pairing prompt text with outputs to compare variations. Prompts can be executed, shared, cloned, and saved to a shared library, supporting collaborative iteration at both the textbook-level and lesson-level.

evaluate how changes impact the content by clicking on buttons that pair each prompt with its corresponding output.

4. **Share:** To fulfill R5 and facilitate collaborative iteration among teams of SMEs, we developed the Shared Prompt Library. This feature allows SMEs to curate effective prompt formulations, enabling others to clone, test, and upvote prompts that perform well during evaluation.

These processes are more cyclical than linear, meaning that sampling, evaluation, and sharing can occur at any stage of the prompt crafting process, allowing for frequent back-and-forth iterations between content team members. Prompt authoring happens at two levels of abstraction:

1. **Textbook Level:** Users create prompts that generate hints applicable across an entire textbook, ensuring broad coverage.

2. **Lesson Level:** Prompts are then refined to address specific lessons, ensuring greater specificity and alignment with particular learning objectives.

In the next two sections, we describe methodological details and results of two studies that we conducted to refine and validate our design: (i) an in-depth, three part study with ten subject-matter experts to refine and validate the interface design and write prompts for generating hints for an entire college-level algebra textbook; (ii) a learning gain study with 358 learners to assess the learning impact of hints generated using PromptHive compared to human-authored hints created previously by SMEs during previous years without AI help. Both studies were approved by the institutional review board.

5.4 Study 1: User Evaluation with Subject Matter Experts

To evaluate how subject matter experts use PromptHive to collaboratively author prompts, we designed a three-part user study involving pre-interviews, collaborative prompt-writing sessions, and post-interviews with experts in math.

5.4.1 Research Questions

We sought to answer the following research questions through this study:

- RQ1: What do SMEs perceive as the most **time-consuming** tasks in manual content authoring, and where can AI offer the greatest assistance through PromptHive?
- RQ2: How do subject-matter experts with prior experience in manual content authoring perceive the **trustworthiness** and **usability** of PromptHive?
- RQ3: Do SMEs feel they can retain **control** over the hint generation process when using PromptHive?
- RQ4: How does the subjective **cognitive workload** of SMEs using PromptHive compare to the manual content generation workflow?
- RQ5: How do prompts **evolve** as subject-matter experts refine them through individual iterations and collaboration with other experts?

5.4.2 Participants

We recruited 10 subject-matter experts (6 women, 4 men) in mathematics tutoring, aged 18-24, who were screened for having prior experience in manually authoring content for OATutor. These participants were either 3rd- or 4th-year undergraduate students or recent graduates, with backgrounds in math-intensive fields such as Data Science, Computer Science, Applied Mathematics, Economics, and Electrical Engineering. All participants except P9 reported having experience tutoring math and other subjects beyond their content authoring work for OATutor. However, P9 had extensive experience authoring content for OATutor, contributing around 500 problems to the platform. Each subject-matter expert was compensated with $150 USD for participating in the study. Table 5.1 summarizes their backgrounds and expertise. Participants had some prior exposure to LLMs, primarily through ChatGPT, but were not advanced users with knowledge of APIs or other technical aspects of prompt engineering.

5.4.3 Procedure & Tasks

In addition to completing a short demographics survey, participants took part in three study sessions, all done only via Zoom video-conferencing:

- **30-minute Pre-Interviews:** These sessions were completed individually and began with participants sharing their prior experience with manually authoring content for OATutor. They were then asked to reflect on this experience and complete the NASA-TLX instrument for measuring perceived workload. Following this, participants were given access to an early

ID	Educational & Teaching Background		OATutor Authoring Experience	
	College Major	**Teaching Experience**	**# of Problems**	**Problem Subject Areas**
P1	Applied Mathematics	TA in Data Science	100	Calculus
P2	Legal Studies, Economics	University tutor for Calculus	150	Physics, Calculus, Algebra
P3	Computer Science	TA for Introductory CS courses	75	Chemistry, Algebra, Physics
P4	Computer Science	High school math tutor, debate coach (middle school - college)	100	Calculus, Statistics
P5	Mathematics	7th-Grade Math Teacher, also tutored AP Statistics and College-level Algebra	150	Algebra
P6	Computer Science, Data Science	Tutor for statistics, math, and programming	200	Statistics, Algebra
P7	Data Science	Tutor for high-school Math	200	Algebra
P8	Computer Science, Cognitive Science	Instructor and TA for Computer Security, Intro to CS tutor	325	Statistics, Algebra
P9	Data Science, Cognitive Science	No tutoring beyond OATutor	500	Physics, Calculus
P10	Electrical Engineering, Computer Science	Tutor for College-level CS	120	Mathematics and Physics

Table 5.1: Educational and teaching backgrounds of participants and their prior experience with authoring hints, including their college majors, teaching experience, the number of problems authored in OATutor, and the subject areas of those problems.

version of the PromptHive interface and asked to complete a series of steps following the 4-stage workflow outlined in Section 5.3.4. During this think-aloud session, the researcher noted any usability issues that emerged, which were then addressed before the collaborative prompt-writing sessions described next.

- **1.5-hour Collaborative Prompt Writing:** These were conducted in groups, with participants choosing between two available time slots. This two-session approach allowed us to simulate both synchronous and asynchronous collaborative scenarios, as prompts written during the first session were made asynchronously available to participants in the second session. The sessions began with participants receiving an overview of how the PromptHive system works, along with a document outlining the necessary instructions and basic prompt-writing strategies adapted from OpenAI's guide to prompt engineering [248]. The supplementary materials from the related publication[1] include a copy of this document to assist future researchers in running similar collaborative prompt-writing sessions using our tool. We excluded technical details and focused on two core strategies: testing changes systematically and writing clear instructions. Participants first worked individually to create and share their best textbook-level

[1] https://dl.acm.org/doi/10.1145/3706598.3714051

prompts, evaluating output quality after testing several variations. These prompts formed the initial set in the shared library, which participants could freely draw from, using them as a base or starting from scratch—during subsequent lesson-level prompt authoring. Next, they were randomly and evenly assigned contiguous groups of lessons (P3 received five lessons, the remaining received six, dividing 59 lessons across 10 participants) to test and refine both their own and others' textbook-level prompts. They could upvote effective prompts and adapt them at the lesson level if the textbook-level versions did not work well as-is. Participants had complete freedom in curating prompts, both in deciding how to frame them and in borrowing or adapting textbook-level prompts shared by others for specific lessons. By the end of the session, each participant contributed a lesson-level prompt to the shared library, collectively covering the entire college-level algebra textbook.

- **30-minute Post-Interviews:** After participants completed the collaborative prompt writing sessions, they scheduled individual post-interviews to share their experience with using PromptHive. They reflected on their experience for generating hints using PromptHive and completed the NASA-TLX [124] instrument for the automated workflow. They also completed the System-Usability Scale (SUS) [197] and the XAI Trust scale adapted from Hoffman et al. [132] questionnaire to share how usable the system felt and whether they trusted using it for the purpose of generating educational content. See Appendix B.1 for the wording of the 8 items in the trust scale. Finally, they participated in a brief semi-structured user interview to unpack their experience with PromptHive.

5.4.4 Materials

We used publicly available Creative Commons Algebra content from Open Adaptive Tutor (OATutor) [262], an open-source adaptive tutoring system based on ITS-principles [261]. This content pool contains materials from the *OpenStax College Algebra 2e textbook* [251]; we contacted the OATutor team to share a copy of the spreadsheet they used during manual authoring. This spreadsheet was ideal because it contained pre-existing human-authored hints and scaffolds that we could compare with those generated using PromptHive, as we do in the second study which focuses on learning gain. We chose mathematics because domain expertise could be particularly valuable for this subject area as LLMs are known to struggle with mathematical reasoning [320], making subject matter expert involvement valuable. It is also an area where responsible applications of generative AI could have immense positive impact on students [226], and recent literature suggests strategies for reducing hallucinations, such as self-consistency [361, 258], an approach that we adopt later for the second study.

5.4.5 Analysis

Our data consisted of interview transcripts and NASA-TLX ratings from the pre- and post-interviews, SUS and XAI Trust ratings from the post-interviews, the textbook-level and lesson-level problems that participants wrote during the collaborative sessions, and the detailed prompt iteration JSON logs captured by the logging engine. We analyzed the qualitative interview data using reflexive thematic analysis [33] through an inductive-deductive lens, using the rich theory on Human-AI collaboration and recent literature on prompt-engineering interface design as a pre-existing code

that guided our interpretations. For the NASA-TLX scores and SUS ratings, we used the standard procedures for calculating scores.

5.5 Study 1 Results

The overall response to PromptHive was largely positive, with participants rating the system as highly usable, achieving an SUS score of 89/100. There was a significant reduction in NASA-TLX ratings for perceived cognitive workload, dropping from 55.17 to 26.73 out of 100 compared to the manual hint authoring workflow (see Figure 5.4). Regarding trust in the AI system, most participants strongly or somewhat agreed that the system felt trustworthy, without causing any wariness, based on the validated measures from the XAI trust scale. See Figure 5.5 for the distribution of responses and Appendix B.1 for the wording of the scale items adapted from [132].

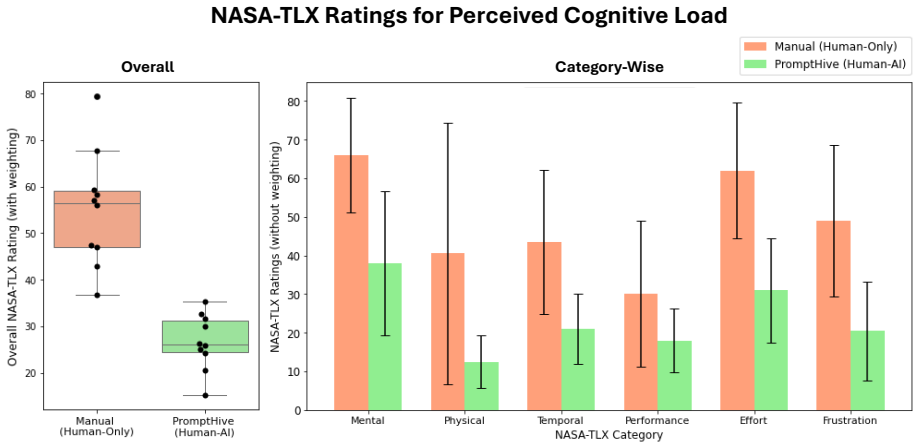


Figure 5.4: NASA-TLX Ratings for Perceived Cognitive Load. The left plot shows overall NASA-TLX ratings (with weighting) comparing manual (human-only) and PromptHive (human-AI) workflows. The right plot provides category-wise NASA-TLX ratings (without weighting) across six dimensions: mental, physical, temporal, performance, effort, and frustration, highlighting significantly reduced cognitive load in the PromptHive condition compared to the manual condition.

5.5.1 Unpacking the AI-assisted Hint Authoring Experience in PromptHive

> *"Honestly, I'm just amazed by how [PromptHive] works. I've spent the last six months making hints, so seeing this in action and knowing that this is what other people will use is pretty incredible. I'm excited to see where it goes."* – P10

During the pre- and post-interviews, participants shared their experience with the PromptHive system and commented on how it compares with the manual workflow and their prior experiences with AI. We grouped the findings from their experiences into six themes (F1-6) described below.

F1: Hint Authoring is Most Time-Consuming and Can Benefit from AI Assistance

To answer RQ1, we asked participants during the pre-interviews which parts of the manual hint authoring workflow they found most time-consuming. We posed this question *before* introducing participants to the PromptHive system to avoid biasing their responses. Multiple participants (P1, P3, P5, P7, P9) pointed to authoring hints (especially those with scaffolded answers), as the most

time consuming, in alignment with our design decision to target that aspect of the content authoring process using PromptHive. For example, P5 shared that, "the most time-consuming part is how to scaffold each step to make it clear and short without including too many words," while P3 mentioned, "each problem has only one statement, but for the hints, there could be 10 or more."

Regarding why scaffolded hint generation was time-consuming, P9 discussed the decision-making process involved in determining the appropriate level of detail for hints, such as whether to explain high-level concepts or break down smaller calculations. They questioned, "If there's a problem that uses the mass-times-density equation, do I need to just say, 'Use a density equation and do the calculation with these inputs,' or go further down to something like, 'What's three times five?"' P10 shared additional considerations when authoring hints, noting that it takes time to "...chain the hints and ensure they flow together."

In contrast, post-interviews revealed a clear consensus among participants that PromptHive significantly reduced the time spent creating hints compared to the manual workflow. Participants noted that they could generate multiple hints and scaffolds in the time it previously took to create just one manually. This efficiency was particularly important when handling large volumes of problems. For instance, P8 remarked, "I was able to create a prompt to generate hints and scaffolds for 15 to 20 questions, whereas before, it would take me at least five, sometimes ten minutes to do just one." P10 added that PromptHive felt "super fast, super intuitive," and noted, "Anyone who's done it manually and then tried the AI system would probably agree that it's much more fun and intuitive... much more enjoyable than just typing it all out."

F2: PromptHive Integrates AI into Expert Workflow More Closely Compared to ChatGPT

Comparing PromptHive to using ChatGPT, P1 mentioned "I also tried, like, ChatGPT, but it always outputs the wrong answer. At least when I was working with this system, the answers were all reliable. I'm curious – how could you generate the result with such a high percentage of correct answers compared to GPT?". Commenting on the tighter integration with the existing expert workflow, P4 noted that PromptHive is "a lot faster because you don't have to type the question out into ChatGPT, and then copy, paste everything onto the spreadsheet. And the formatting for the spreadsheet was also like, difficult to do, because in it you had to write six or seven rows for each question. So I think the system is a lot easier to use. I would recommend over manually putting things into ChatGPT."

F3: Participants felt that the AI-assisted workflow in PromptHive led to more consistent hints

Several participants (P1, P7, P8 and P9) felt that the AI hints were more consistent than the human-only hints because of the variability in style or structure of the hints across different subject matter experts. Sharing reasons for the variability they observed, P1 emphasized that different members had different teaching styles. P7 felt the AI workflow was more uniform and less prone to error compared to the manual workflow. Pointing to a different reason for the observed inconsistency, P4 mentioned that the quality of hints would vary a lot, and "sometimes you could see the effort decreasing from like the first question to like...so, yeah, I think it would vary, yeah", as people got tired.

F4: Participants felt that they mostly retained control over Hint Generation

Many participants (P2, 6, 8, 9, 10) reported that they strongly felt they could steer the output

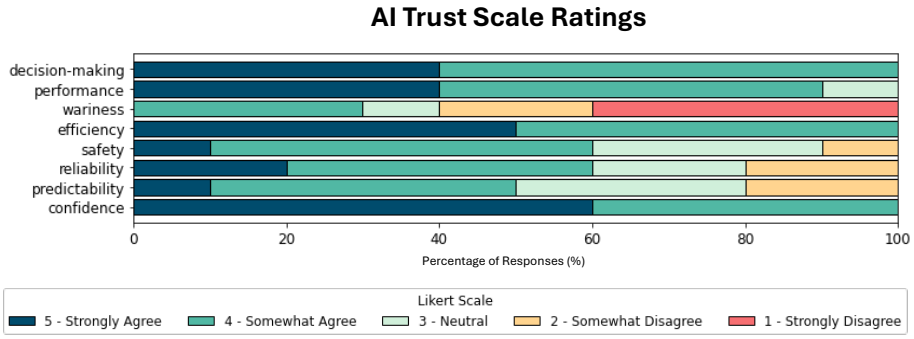


Figure 5.5: AI Trust Scale Ratings. The stacked bar chart shows the distribution of participants' responses across eight dimensions of trust in AI: decision-making, performance, wariness, efficiency, safety, reliability, predictability, and confidence. Note that for wariness, higher *disagreement* is desirable. For other items, higher agreement is desirable.

of the LLM and apply their prior subject matter expertise in authoring prompts and evaluating the model output. For example, P6 was able to make hints simpler or complex by altering the student age: "I tried asking the AI to explain to a 7-year-old, and the output was simple. When I changed it to a 10-year-old, it became more complex." P6 removed redundancy by instructing the LLM to not repeat hints and limit the number of scaffolded hints, like P9. P8 mentioned that they could eventually get the AI to follow instructions after iterating on the prompt a few times: "even if there are times where they are acting a little bit, you know, a little bit weird, like with capitalization issues or the tone is not right, I can just instantly update it. And then, usually, after two or three tries, I can get like, a version that I think would be even better than something that I would write myself."

Other participants felt less strongly about their ability to control the AI output, e.g., P4 mentioned that while the AI model felt "steerable to an extent", it could sometimes be "stubborn" and difficult to steer for general textbook level prompts but that it felt easier to steer prompts at the lesson level. P5 felt that sometimes they struggled to get the AI model to adhere to multiple opposing instructions, especially when the prompts were complex, e.g., in trying to simplify the hints for a younger audience, the model would inadvertently give responses that felt too long.

F5: Subject Matter Experts did not feel replaced by the AI workflow

Reflecting on the ongoing discussions surrounding how generative AI will affect education, P8 said "I think when I hear AI and LLMs and education right now, I think of a talk that I just attended that was just very extremely pro-LLM. And I think we do need to be a little careful about how we inject LLMs into education. But I think having this idea where the LLM isn't creating the questions, the questions are already there. That is good to me. But then also, beyond that, like your job is to generate prompts...at the end of the day, having a prompt engineer who still goes through the process of needing to understand how to guide a student, I think that that still can be very beneficial, versus just like a solely AI based thing." When asked if they felt replaced by the AI model, they said "I don't think the AI system can replace me. I think my role in terms of telling the AI what to do, that involvement is really important." P6 felt that more complex problems could benefit from increased oversight of subject matter experts.

F6: Collaborative authoring can help Subject Matter Experts be more creative with

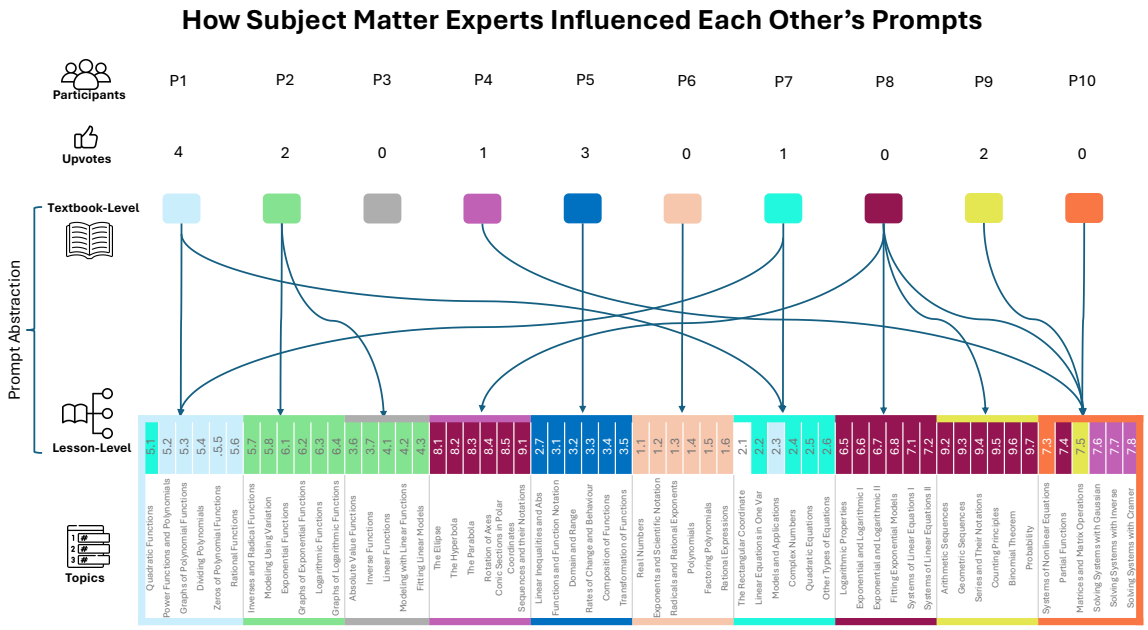


Figure 5.6: Participants' influence on each other's prompts. Outer border colors for lessons correspond to lessons assigned to participants, with the same color representing the same participant. The fill color of lesson numbers and the arrows indicate the textbook-level source for the lesson prompts.

prompts

The collaborative aspect of the AI assisted workflow in PromptHive was well-received by many participants, who felt that sharing ideas led to a richer variety of more creative prompts. P10 said "seeing other people's prompts gave me ideas that I didn't even think to ask for" and P8 felt "the collaborative aspect was wonderful. Seeing how others structured their prompts helped me brainstorm and improve my own." Comments from P5 indicated that exposure to prompts from others helped them think beyond their own experience as a middle school teacher – "I remember seeing prompts asked for hints to have a positive tone, stuff like that. I think it inspired me to think about that because currently I'm a middle school teacher, so I only think about making things short so people will look at it. When I saw another prompt which said you are a tutor in college, you are a professor, that made me think about the main population of students this model is targeting." Figure 5.6 provides details on how participants influenced each others' lesson-level prompts.

5.5.2 Exploring How Subject Matter Experts Iterate on Prompts

The PromptHive logging engine captured data on user interactions whenever subject matter experts executed a prompt or saved it in the shared library. Figure 5.9 and 5.8 shows the distribution of executions and saves across the ten subject matter experts, indicating that on average, they made 30 executions ($SD = 12.57$) and 10 saves each ($SD = 5.02$) to author the 10 textbook-level prompts and 59 lesson-level prompts covering hints for the entirety of a textbook. The logging engine also links the prompts captured into a tree structure that can be exported as a JSON file, widely used and lightweight data-interchange format that is easy for humans to read and write [62]. See Figure

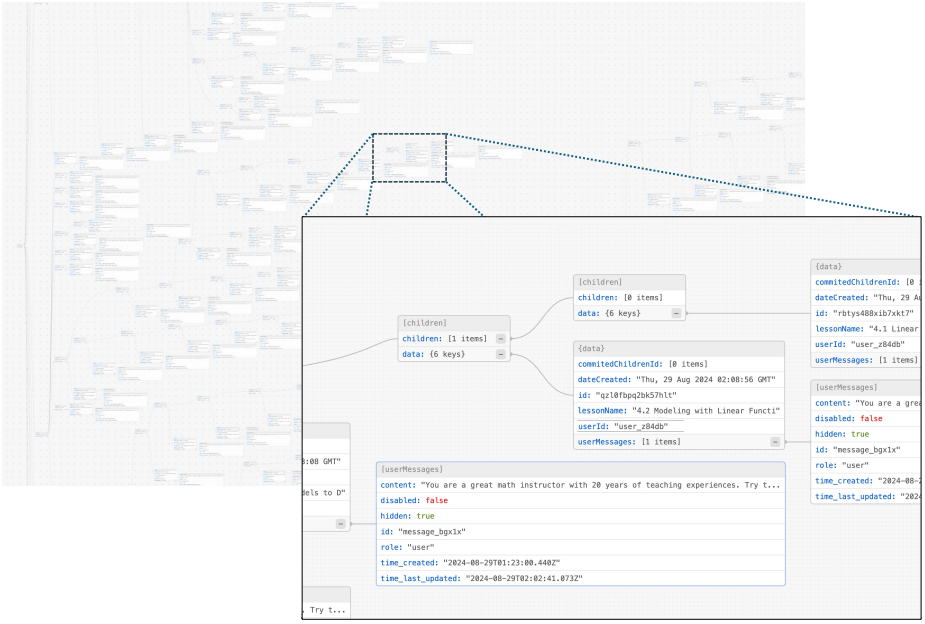


Figure 5.7: A snapshot of the data captured by PromptHive's logging engine, illustrating how researchers can retrace the iterative process of domain experts when refining prompts. Each {`data`} node in this example represents an execution of a prompt variation within the scratchpad, and the linked {`userMessage`} node contains the prompt text. A similar logging mechanism tracks how prompts are saved to the shared prompt library.

5.7 for a snapshot of the rich interactions captured during the 90 minute sessions.

From this log, we found that participants did between 1 (P2) and 17 (P3 and P8) iterations based on executions to arrive at their final textbook-level prompt, with most taking between 3 to 6 iterations. Below is an example of the textbook-level prompt chain from P5, with the removals and additions across each iteration highlighted in red and green, respectively:

- **Iteration 1:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of ELD students who have ADHD and severe learning disability to understand math. Make sure the hints that you give are concise which means less than 10 words for each step. They are also easy to understand, encouraging, and interesting for students to follow. Those students are also historically known to be marginalized.

- **Iteration 2:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of ELD students who have ADHD and severe learning disability to understand math. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow. Make sure the hints that you give are concise which means less than 10 words for each step. They are also historically known to be marginalized.

- **Iteration 3:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents.

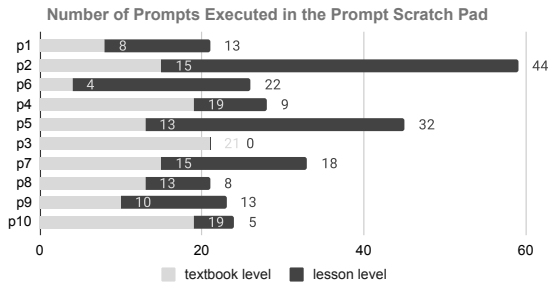


Figure 5.8: Number of prompts **executed** in the Prompt Scratch Pad by each participant, categorized into textbook-level and lesson-level prompts. This chart illustrates the level of engagement by participants during the prompt authoring process.

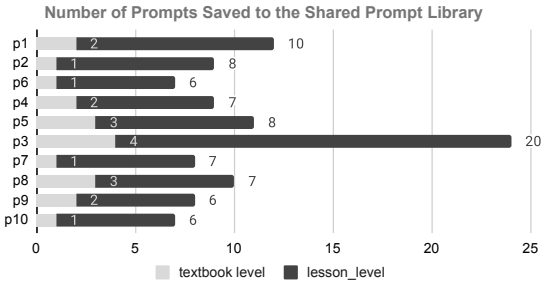


Figure 5.9: Number of prompts **saved** to the shared Prompt Library by each participant, categorized into textbook-level and lesson-level prompts. This chart highlights participants' contributions to the shared resource after refinement.

> Currently, you are working with a group of ELD students who have ADHD and severe learning disability to understand math. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow. No change from Iteration 2.

- **Iteration 4:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of ELD students who have ADHD and severe learning disability to understand math. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow. No change from Iteration 3.

- **Iteration 5:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of special education students. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow. ELD students who have ADHD and severe learning disability to understand math. **[sic]**

- **Final Prompt:** You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of special education students. You need to add some emojis to each hint to make it interesting for students to follow. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow.

Here, we see how P5 starts off by experimenting with writing hints for students with learning disabilities, even though the focus of the textbook was for college-level math, likely drawing from their experience in teaching diverse students as a 7th grade math teacher. In iteration two, they remove the last two sentences on making the hints concise and being mindful of the marginalization of such students. Given how LLMs can generate different outputs for executions of the same prompt, we see there was no change to the content from Iteration 3 to 4. In 5, they mention special education twice, likely to emphasize who they want these hints to target. Finally, they add a note about including emojis to make the hints more engaging. The JSON log captures similar chains for both textbook-level and lesson-level prompts across all participants to offer empirical insights on how

subject matter experts iterate prompts. See Appendix B.3 for a list of finalized textbook level prompts from all participants.

Looking at how participants' textbook-level prompts influenced each others' lesson-level prompts, as shown in Figure 5.6, we find that upvotes did not correlate with influence, as the most influential post from P8, which influenced 4 other participants, had 0 upvotes. Apart from P3's prompt, which did not influence any lesson-level prompt, most participants influenced at least one group of lesson-level prompts (often their own), and sometimes those of several others, in alignment with user comments on collaboration being helpful in prompt-authoring. Of the 59 lesson-level prompts committed to PromptHive, only 8 were verbatim textbook-level clones, i.e., they didn't need tailoring. The domain experts chose to tailor the remaining 51.

5.6 Study 2: Learning Gain Study with College-Level Math Learners

In this study, we explore the impact of PromptHive hints on learning gains compared to a human-only authored hints control condition.

5.6.1 Research Questions

- RQ5: Do PromptHive-generated hints lead to learning gains, and how do those gains compare to those from human-only hints?

5.6.2 Participants

Through Prolific (a crowdsourcing platform), we recruited a total of 358 current undergraduate college students in the United States. As the recruitment was conducted through Prolific, the sample may not fully represent the broader undergraduate population in the US. To address this, we intentionally kept recruitment criteria broad to include students from diverse academic backgrounds (e.g., STEM, Humanities) and various age groups to enhance generalizability. Each participant was compensated with $20 USD for study completion.

Since participants completed a sequence of 3 lessons, we excluded any lesson submissions where participants did not fully complete all parts of the lesson sequence (i.e. 3-question pre-test, 5-question hint condition, 3-question post-test). After this exclusion, we resulted in 225 unique participants, with a total of 549 completed lesson submissions (268 for human-only and 281 for PromptHive).

5.6.3 Tasks

Participants completed a 3-question pre-test to assess their initial understanding of the lesson's topic. They were then presented with 5 additional questions, receiving correctness feedback on their answers along with a tutoring pathway based on the assigned condition (either a human-curated hint pathway or the PromptHive-generated pathway). Finally, participants completed a post-test with the same 3 questions as the pre-test to assess learning gains. This sequence was repeated for 3 distinct lesson/condition pairings. At the end of the experiment, participants were shown a survey code and asked to enter it into their Prolific portal.

5.6.4 Materials

For the human-only control hints, we use the hints generated by the OATutor project, contained within the OATutor system. These hints were multi-level hint pathways, curated by subject matter experts with prior tutoring experience. There was no restriction on the number of hints and scaffolds to incorporate in the hint pathway.

The prompts utilized for PromptHive to generate hint pathways for all questions in a single response were those created by the content team members from study 1. These prompts specified detailed instructions for rendering problems correctly in the OATutor system, with mathematical expressions following the required syntax for proper display and functionality. Each response generated by PromptHive was a JSON object with 80 keys, one for each question. The value for each key was a string that, when parsed, contained a multi-hint pathway specific to that question. Building on the methodology of a study that employs self-consistency to reduce hallucinations for worked solutions in mathematics, achieving near 0% hallucination in College Algebra [361, 258], we used a similar approach for hallucination mitigation. Since each response string by PromptHive contained a multi-hint pathway rather than a worked solution, we used a similar approach to hallucination mitigation, but one which involved vectorizing all the responses to a single question and finding the most representative response (closest to the centroid) [199]. In essence, we prompted PromptHive to generate this JSON object 30 times, resulting in 30 versions of the multi-hint pathways for each of the 80 questions. For each set of 30 responses for a single question, we first vectorized the responses using SentenceTransformer and found the centroid vector. Then, we used cosine similarity to find the response closest to the centroid vector, which represents the most representative response out of our 30. This response was then used in our study and put into the OATutor system as being the multi-hint pathway for the PromptHive condition.

In order to ensure compatibility with OATutor, we checked that each of the 80 questions was able to be rendered correctly in the system, complying with its formatting guidelines. This check resulted in having to make slight modifications to the PromptHive generated responses, such as changing questions with multiple correct answers to "multiple choice" and "string" (exact-match) answer types. These modifications were made by the second and third authors of the paper, taking them 3 hours and 15 minutes and 6 hours and 19 minutes to complete, respectively. Sample hint pathways for both the PromptHive hints and human-only control hints are shown in Appendix B.2.

5.6.5 Procedure

We used the Qualtrics platform to facilitate random assignment of learners to either the human-only authored hints control or the PromptHive experiment condition. Prolific was used to recruit participants. All participants first viewed an instructions screen which explained the task, after which they were randomly assigned to one of twenty lesson/condition configurations (10 lessons, 2 conditions).

The OATutor system's log data was used to track user actions, including inputting answers, opening hints, conditions assigned, timestamps, and other interactions.

5.6.6 Analysis

We assessed the normality of average pre-test scores, post-test scores, learning gains, and time-on-task using the Shapiro-Wilk Test of Normality. Since we rejected the null hypothesis (indicating that the data was not normally distributed), we used the Kruskal-Wallis test to analyze pre-test scores and time-on-task to assess evenness at pre-test and detect statistically significant differences in time spent per lesson. Subsequently, we utilized the Mann-Whitney U test for pairwise comparisons. If we had failed to reject the null hypothesis of the Shapiro-Wilk Test of Normality (indicating normality), we would conduct the same analysis using ANOVA.

To examine pre- to post-test learning gains for the lesson and condition pairings, we used the Wilcoxon signed-rank test for paired samples since the data was not normally distributed. Otherwise, we would have utilized a paired t-test for the same comparisons. Finally, for understanding how the hint conditions compare to one another, we utilized a mixed linear model (with ranked data due to non-normality), incorporating fixed effects of the condition and random effects for the participant. Specifically, the following mixed-effects model was used in the analysis:

$$\text{Learning Gain}_i = \beta_0 + \beta_1 \cdot \text{condition}_i + u_j$$

where β_0 is the intercept, β_1 is the fixed effect of the condition, u_j is the random effect for participant j, and ϵ_i is the error term.

5.6.7 Results

The Shapiro-Wilk Test indicated non-normality for average pre-test scores, post-test scores, learning gains, and time-on-task. Therefore, we further utilized Kruskal-Wallis and failed to reject the null for all lessons and thus found both hinting conditions to be even at pre-test for each lesson, allowing for sound comparisons. Average pre-test scores were 51.9% for the PromptHive hints condition and 45.9% for the human-only control. When comparing time-on-task between the two hint types, the Kruskal-Wallis test revealed statistically significant differences overall (p = 0.031). Through the Mann-Whitney U test, we found that only lesson 8.4 showed statistically significant differences between time-on-task of the PromptHive hints and human-only authored hints (p = 0.030).

Next, we utilized the Wilcoxon signed-rank test for paired samples and found both the PromptHive (p ¡ 0.001) and the control (p ¡ 0.001) conditions to exhibit statistically significant differences, indicating that both conditions facilitate learning. Table 5.2 shows these results granularly at the lesson level, with 4 lessons showing statistically significant learning gain with PromptHive compared to 2 lessons for the human-only authored hints. Average learning gain in the control condition was 7.47% and for PromptHive was 8.13%, with 3 lesson/condition pairings exhibiting negative learning gains (4.3 PromptHive, 4.3 control, and 8.4 control). Figure 5.10 shows these learning gains. Due to non-normality, we utilized the mixed linear model described previously with ranked data. We found no statistically significant differences between the learning gains of the hint conditions (p = 0.688).

5.7 Discussion

Our findings demonstrate how PromptHive can integrate subject matter experts into the prompt engineering process, effectively combining human expertise with the efficiency of generative AI.

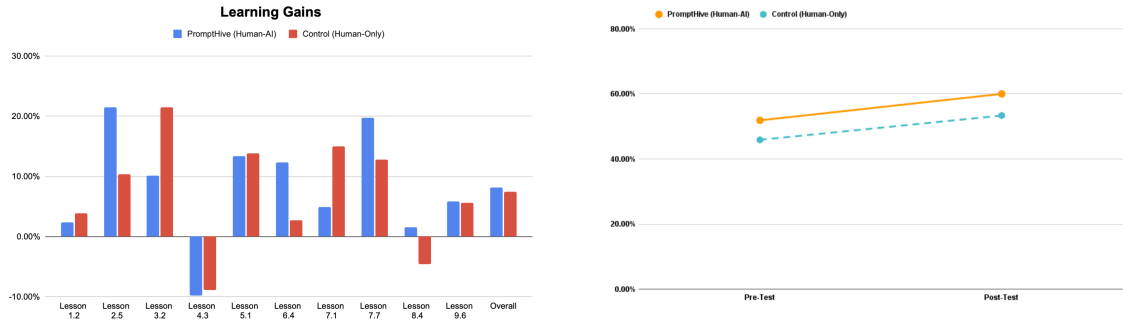


Figure 5.10: Learning gains by lesson (left) and overall pre- to post-test scores (right). The lesson-level bar chart illustrates generally positive learning gains across lessons, with comparable performance between hints authored manually and those created using PromptHive. The pre- to post-test learning curves (right) demonstrate consistent learning gains, with both PromptHive and control groups showing comparable gradients, indicating that PromptHive can match manually authored content from experts.

Through two human studies – first with experts who used PromptHive to produce hints for an entire college-level algebra textbook in OATutor, an adaptive tutoring system, then with a general learner population to assess how the learning gain compares with manually authored content – we showed how content produced using PromptHive can match the quality of manually-authored content while significantly reducing the overall time and effort required, without compromising expert oversight, sense of control, or trust. Building on these results, we now reflect on design recommendations for human-LLM collaboration, exploring opportunities for applying PromptHive beyond mathematics and OATutor.

5.7.1 Increasing Automation while Retaining Human Control

In alignment with Shneiderman's optimistic vision of human-centered AI automation [314], where the goal is to design AI systems that increase automation while retaining human control, our findings offer solace to HCI and education researchers who are concerned about generative AI replacing human experts. In PromptHive, we bring subject matter experts back to the forefront of AI-assisted educational content production by elevating their role from content producers to curators, involving them deeply in the prompt authoring process in a way that preserves their agency over the substance and style of the hints.This places PromptHive in a unique position among similar systems, as other tools have overlooked SME agency for co-designing with generative AI [273]. PromptHive's unique A/B testing framework, and its facilitation of collaboration between SMEs, through prompt sharing and collaborative authoring, offer unique advantages in usability and scalability. Our findings also indicate the value of bringing together multiple human experts and helping them learn from each other when working with powerful, albeit unpredictable tools like LLMs.

Design Recommendations

To retain human control, designers of future AI-assisted content production tools should strive to provide explicit interface support for quickly viewing AI-generated content in context of related information and mechanisms for comparing the impact of changes across different prompt formula-

Lesson	Condition	N	Avg. Time (s)	Learning Gain (%)	Avg. Pre-test (%)	Avg. Post-test (%)	p-value
1.2	PromptHive	29	480.0	2.31%	74.71%	77.02%	0.315
	Control	26	585.5	3.85%	65.38%	69.23%	0.224
2.5	PromptHive	31	663.0	21.50%	47.31%	68.81%	**0.027**
	Control	29	533.0	10.34%	37.93%	48.28%	0.235
3.2	PromptHive	33	449.0	10.11%	50.51%	60.62%	**0.024**
	Control	28	312.0	21.44%	41.66%	63.11%	**0.001**
4.3	PromptHive	34	579.0	-9.80%	80.39%	70.59%	**0.026**
	Control	30	459.5	-8.88%	74.45%	65.56%	0.225
5.1	PromptHive	25	1380.0	13.34%	42.66%	56.00%	0.147
	Control	29	748.0	13.80%	34.47%	48.27%	0.068
6.4	PromptHive	31	667.0	12.37%	54.30%	66.67%	0.159
	Control	25	444.0	2.67%	56.00%	58.67%	0.725
7.1	PromptHive	27	755.0	4.95%	66.66%	71.61%	0.302
	Control	29	1025.0	14.93%	63.23%	78.16%	**0.025**
7.7	PromptHive	27	1325.0	19.76%	27.15%	46.91%	**0.015**
	Control	26	1225.0	12.83%	30.76%	43.59%	0.112
8.4	PromptHive	21	880.0	1.60%	52.38%	53.98%	0.753
	Control	22	624.0	-4.55%	43.94%	39.39%	0.418
9.6	PromptHive	23	1566.0	5.79%	7.24%	13.03%	0.102
	Control	24	1430.5	5.55%	4.16%	9.71%	0.102

Table 5.2: Learning gain results comparing PromptHive (Human-AI) and Control (Human-Only) conditions across lessons. Significant differences (p ¡ 0.05) in pre- to post-test scores are bolded.

tions. The A/B buttons in the PromptHive Scratchpad that pair prompt variations with generated output exemplify this approach. Fostering collaboration between multiple human experts can also play a role in maintaining human oversight and trust and accelerate adaptation of general prompts to specific contexts. In PromptHive, we saw how the Shared Prompt Library allowed subject matter experts to share their best prompts, observe and try out peer approaches to prompt engineering, and explore variations to identify ways to adapt prompts to particular lessons.

5.7.2 Increasing Visibility into the Prompt Authoring Process

PromptHive's logging engine captured rich interaction data on how experts iterated on prompts, adding empirical richness and providing granular insights that would have been difficult to obtain through interviews alone. For instance, we observed how P5 experimented with prompts aimed at younger students with special needs before finalizing a prompt for a college-level audience. Their focus on tone, including the use of encouraging language and emojis, resonated with many other participants, who remarked on these features during post-interviews. We also learned that upvotes did not correlate with influence, as evidenced by the fact that four other experts adopted P8's prompt, despite it having 0 upvotes.

The logging engine enables researchers to move beyond viewing prompt writing as a static process focused solely on the final formulation. Instead, it allows them to study it as a dynamic process, tracing the evolution of prompt formulations and their corresponding outputs through a series of interim stages — from an expert's initial conception of a prompt, to when they deem it ready to share, and finally to how others continue to adapt and incorporate it into their own prompts. This dynamic view is better aligned with the practical realities of generative AI research, where prompts

written today are likely to become obsolete as newer models emerge. PromptHive's logging engine equips researchers to better understand and address this ongoing evolution.

Design Recommendations

Since PromptHive is fully open-source, future HCI researchers and designers can leverage its logging engine to study how multiple experts collaboratively iterate on prompts. Our analysis of the logged data revealed that surface-level quantitative metrics, such as upvotes, can provide an incomplete picture of true user preferences. Therefore, designers should complement quantitative metrics with insights derived from qualitative data trails to gain a more comprehensive understanding of how users refine prompts to achieve desired model outputs. Designers could also use the logging engine to create visualizations that provide experts with insights into the authoring process. For instance, if a subject matter expert encounters a 'dead end' where a particular formulation refuses to work, they could examine qualitative data trails from their own attempts or those of others to identify alternative approaches that might be more effective.

5.7.3 Applying PromptHive to Different Domains and Systems

Applying PromptHive to Subjects Beyond Math

We chose to focus our evaluation in a math education context given the immense opportunities [258, 370], challenges [231, 324], as well as real-world interest from educators in applying AI in this context [159, 235]. However, given that PromptHive seamlessly integrates with OATutor [261], which features a content repository that is being expanded to other subject areas including Chemistry, Physics, and Data Science [247], the current implementation of PromptHive can easily be adapted to supporting those areas. For example, since conducting our studies, PromptHive has been successfully piloted with Chemistry instructors. Such extensions are also possible because of the generalizability and versatility of the underlying foundation models, which are capable of producing content across many subject areas.

Integrating PromptHive with Systems Beyond OATutor

PromptHive interfaces with OATutor via the publicly available Google Sheets API for reading and altering its underlying structured data source, and outputs content using the widely used JSON format, a lightweight data-interchange format that is easy for humans to read and write. The specific output attributes are outlined in a system-level prompt that can be adapted to other structured data sources. PromptHive features an advanced prompt authoring mode that exposes the system-level prompts so that users can alter the structure of its output to different systems beyond OATutor.

Furthermore, the four-part workflow in PromptHive - *loading* structured content (initially from an existing human-only expert workflow), *authoring* prompts that generate AI content integrated with the structured content, giving experts an opportunity to view the output in context, *iterating* on it through trying out alternative formulations, and *sharing* the most effective formulations with other subject matter experts so they can borrow ideas from them and refine them further – embodies elements that aren't unique to OATutor. This workflow contributes to existing literature in HCI highlighting the value of experimentation in guiding LLM output and offers a concrete characterization of how to design human-centered prompt-authoring tools for augmenting structured data

sources. For example, it could be applied to other contexts such as the development of educational curricula. Given the outlines of individual lessons, the four-part workflow could potentially serve as a method of rapidly generating class curriculum, lending itself to the pre-existing well-defined nature of the content's structure.

Extending PromptHive to Contexts Beyond Education:

The PromptHive workflow could conceivably be abstracted to any scenario involving text production micro-tasks that require consideration of contextual information and structured output. In these scenarios, there is value in migrating human labour away from this type of tedious work to instead authoring and maintaining policies that effectively generalize across related contexts. In such scenarios, this workflow promotes the worker from a doer role to a managerial role, embodying the producer-to-curator shift.

This would not include long form writing tasks, such as authoring a novel, but could include writing up expense reports or responding to customer support emails [11], where representatives become prompt engineers focused on authoring general prompts (i.e., "textbook-level prompts") to respond to customer queries, as well as tailoring prompts for dealing with particular sub-categories of support issues (i.e., "lesson-level" prompts). Much like our scenario where we had a collection of manually authored hints and participants with experience writing them, customer service reps could use existing emails they've handled to evaluate how PromptHive-generated responses compare in customer satisfaction or problem resolution scores, similar to our learning gain scores.

Design Recommendations

Designers seeking to integrate rapid collaborative iteration on prompts by subject matter experts can apply PromptHive's four-part workflow to different contexts where AI supports the augmentation of a structured data source. In doing so, they should prioritize adapting to the needs and nuances of the expert workflow rather than requiring experts to conform to the limitations of the LLM. This approach contrasts with many AI integrations, where the LLM functions as a side tool, processing unstructured input and generating unstructured output.

5.7.4 Limitations and Future Work

There are 4 main limitations in our research. Firstly, the quality of the output from PromptHive is inherently tied to the capabilities of the underlying foundation models, which carry significant risks including bias, hallucinations, and a lack of transparency. Awareness of these risks is crucial when deploying PromptHive in large-scale educational scenarios. While mitigating these risks is beyond the scope of this work, our research offers a pathway to address them by increasing participation and oversight from human experts. However, when applying AI systems like PromptHive to large and diverse student populations, it is essential to exercise caution and ensure these risks are appropriately managed.

Secondly, the participant group for our first study consisted of senior undergraduate students or recent graduates with prior content-authoring experience. While these participants were selected for their teaching experience across STEM subjects (e.g., mathematics, physics, chemistry, data science, computer science, and statistics), their limited subject matter expertise and narrower perspectives

may affect the generalizability of our findings. Future research should include experts from more diverse fields, particularly non-STEM domains, to explore challenges unique to these areas.

Thirdly, the generated content is not directly editable by SMEs in the PromptHive interface. We intentionally limited direct editing to encourage participants to focus on improving the overall quality of outputs by refining the prompts themselves, rather than editing the generated content to bypass issues with the prompts. Future iterations could incorporate the ability to directly edit outputs as well as provide SMEs with assurance that the hints they see are identical to what students will see (for example, by setting the default temperature to zero).

Finally, we chose not to involve subject matter experts in system-level prompt design to maintain their focus on pedagogy. However, when prompts become too complex, involving subject matter experts in system-level prompting could prove beneficial. This may require additional scaffolding or support from AI experts, extending the PromptHive workflow to include collaboration between domain experts and AI experts. Moreover, including users in system-level prompt design could help reduce the need for manual corrections to rendered outputs and address periodic syntax issues in some of the generated hint pathways. Around 40% of the total manual time spent producing hints was spent by editors on correcting relatively minor syntactic issues in the output of PromptHive. Future design revisions could focus on involving editors in system-level prompt iterations to mitigate these rendering and syntax issues, rather than concentrating solely on pedagogy.

5.8 Conclusion

We designed PromptHive in collaboration with ten mathematics tutoring subject matter experts to allow them to author prompts, immediately see LLM outputs across a number of different problem contexts, and then save these outputs within the native publishing workflow of an adaptive tutoring system project. Our study of the subject matter experts found that PromptHive cut NASA-TLX subjective cognitive load in half, from 55.17 to 26.73 and reduced the total time to author hints by factor of 30. This reduction in workload and increase in efficiency was accomplished without a decrease in content quality, as measured by our randomized controlled study involving 358 learners and 10 lessons showing significant learning gains from hints produced with PromptHive (8.13%, p ¡ 0.001) that were not statistically significantly different from the learning produced by the more laborious and time consuming manual hint authoring process (7.47%) that did not utilize PromptHive ($p = 0.688$).

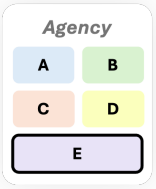
Participants remarked on how much they valued the collaborative aspect of the system, learning from each other's prompt ideas while retaining editorial control of the produced content. This interest in prompt improvement and retaining their expert judiciousness was further made evident when participants were given the option to use one of the existing shared textbook-level prompts verbatim, but in 86% of the lessons instead chose to further tailor a prompt to the lesson. These positive user experiences were reflected in an "Excellent" overall average system usability score of 89/100. PromptHive demonstrates a way forward for human-controlled AI content production that migrates workers from strictly producers to curators and shows how, with effective design, LLMs can improve task efficiency as well as domain expert satisfaction.

5.9 Acknowledgements

We thank the Vice Provost of Undergraduate Education's Micro Grant Program at UC Berkeley and the 2023-2024 Tools Competition for providing financial support for this work. We also extend our gratitude to Joe Fang and Sarva Sanjay from the Department of Computer Science at the University of Toronto for their contributions, with Joe assisting in deploying the prototype and Sarva implementing the back-end logging engine.

Chapter 6

Preserving Human Agency



> It's very important right now for people to be working on the issue of how will we keep control. We need to put a lot of research effort into it.
>
> *Geoffrey Hinton [131]*

Research Context: In this chapter, we move from discussing specific systems like ABScribe and PromptHive to providing a broader understanding of how to integrate AI into content authoring while preserving human agency. The systems from prior chapters serve as concrete examples of how to integate AI into content authoring, but, as alluded by Hinton's quote [131] at the beginning, we need to put a lot of research effort into understanding how humans can retain control as technology increasingly permeates various aspects of the authoring process. To that end, we delve deeper on the third and fourth primary research questions through a systematic review and an interview study involving 1676 HCI papers and 15 authors. We synthesize strategies for AI-assisted writing support from over 100 proposed and existing systems in recent HCI literature (including ABScribe) into four overarching design strategies that future designers of AI systems can adopt.

Related Publication:

Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. 2025. *Co-Writing with AI, on Human Terms: Aligning Research with User Demands Across the Writing Process.* Manuscript under review. Preprint available at https://arxiv.org/abs/2504.12488

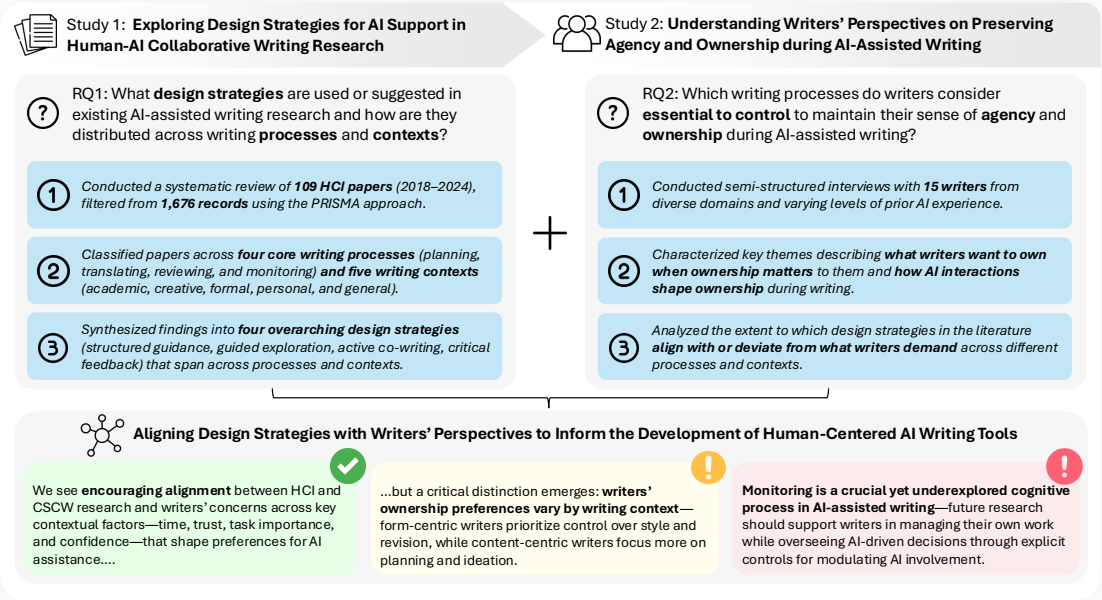


Figure 6.1: We present two inter-connected qualitative studies exploring **how to design for human agency in Human-AI Collaborative writing**: **(1) Study 1:** A systematic review and thematic analysis of 109 papers (2018-2024), carefully selected from over 1,600 in the Human-AI collaborative writing literature using the PRISMA methodology; **(2) Study 2:** A semi-structured interview study with 15 writers, each bringing diverse experience across writing genres, varying familiarity with AI tools, and knowledge of generative AI.

Abstract: As generative AI tools like ChatGPT become integral to everyday writing, critical questions arise about how to preserve writers' sense of agency and ownership when using these tools. Yet, a systematic understanding of how AI assistance affects different aspects of the writing process–and how this shapes writers' agency–remains underexplored. To address this gap, we conducted a systematic review of 109 HCI papers using the PRISMA approach. From this literature, we identify four overarching design strategies for AI writing support–*structured guidance*, *guided exploration*, *active co-writing*, and *critical feedback*–mapped across the four key cognitive processes in writing: *planning*, *translating*, *reviewing*, and *monitoring*. We complement this analysis with interviews of 15 writers across diverse domains. Our findings reveal that writers' desired levels of AI intervention vary across the writing process: content-focused writers (e.g., academics) prioritize ownership during *planning*, while form-focused writers (e.g., creatives) value control over *translating* and *reviewing*. Writers' preferences are also shaped by contextual goals, values, and notions of originality and authorship. By examining when ownership matters, what writers want to own, and how AI interactions shape agency, we surface both alignment and gaps between research and user needs. Our findings offer actionable design guidance for developing human-centered writing tools for co-writing with AI, on human terms.

6.1 Introduction

As Large Language Models (LLMs) grow more powerful and pervasive, AI tools like ChatGPT have become integral to the modern writing process. Computers are evolving from mere *tools* to collaborative *companions*, raising concerns about the encroachment of writers' co-creative boundaries [25], eroding their agency and ownership over the writing process [193, 192]. At the same time, AI tools have proven to be remarkably helpful to writers, augmenting their capabilities across the composition gamut ranging from brainstorming and ideation [302, 317], to editing and revision [282, 187]–and everything in-between [227]–making it impractical to abandon this technology altogether.

One way to address the tension between ensuring human control and increasing automation [315] is to examine how existing and proposed AI-assisted writing systems in the Human-AI collaborative writing literature support distinct cognitive processes in writing [94], and whether that support encroaches on what writers consider to be central to preserving their sense of agency (i.e., their perception of control and autonomy over the writing process), ownership (i.e., their feeling of personal investment in and attribution of the final text), and task delegation (i.e., their choice about which writing subtasks to assign to AI, based on which cognitive processes they consider essential). Writers must maintain agency over the cognitive processes they value most through careful task delegation in order to preserve their sense of ownership over the final text. However, despite the unprecedented pace at which the AI-assisted writing research has grown over the past few years, there is currently no systematic understanding of what cognitive processes are being supported in numerous AI-assisted writing tools, what strategies are being used to offer that support, and how those strategies align with user perspectives across different forms of writing. Furthermore, these strategies have largely been evaluated from a usability and efficiency perspective [26], treating writing as a single task centered on optimization, with limited attention to preserving human agency through effective human-AI collaboration [6] and across its distinct cognitive processes [94].

In this work, we explore whether recent research on Human-AI collaborative writing aligns with user needs, shifting the focus away from output-oriented concerns like productivity toward human-centered considerations–particularly how to preserve writers' sense of agency and ownership when collaborating with AI. To investigate this, we ask two research questions:

- **RQ1:** What design strategies are used or suggested in existing AI-assisted writing research, particularly in terms of interaction models and the intended use of AI outputs, and how are these strategies distributed across writing processes and writing contexts?
- **RQ2:** Which cognitive processes do writers consider essential to control in order to maintain their sense of agency during AI-assisted writing, and how do user situations, writing contexts, and AI interaction types shape their perceptions of ownership?

To answer RQ1, we conducted a systematic review and meta-analysis of 1,676 papers in the Human-AI collaborative writing space from 2018-2024. We first analyze the systems developed or proposed in recent Human-AI collaborative writing literature, then classify them according to distinct thinking processes involved during composition, as outlined in Flower and Hayes cognitive process theory of writing [94]. Then, to answer RQ2, we interviewed 15 writers across diverse domains, exploring how AI affects their sense of control and creative ownership across different writing processes. We then synthesized findings from both studies, revealing encouraging alignments

as well as notable gaps between current AI writing support systems and writers' expressed needs. Figure 6.1 provides an overview of our research approach.

Our synthesis offers actionable guidance for designers, highlighting specific areas and methods of AI support that prioritize user agency. Rather than supporting all aspects of writing indiscriminately, our work helps focus design efforts on features that meaningfully preserve writers' sense of control and ownership. Grounded in these findings, our contributions to the CSCW Human-Centered AI community include:

- The first comprehensive study on designing for human agency in AI-assisted writing, combining a systematic review of post-generative AI research with a user-centered analysis of how writers seek to preserve ownership and originality.
- A detailed characterization of four overarching design strategies for AI writing support, grounded in writers' perspectives on when ownership matters, what they want to own, and how AI interactions shape that ownership.
- Actionable design guidance for CSCW and HCI researchers developing AI writing tools, including concrete recommendations for supporting writer agency across the cognitive processes of writing. Our work informs future systems that foreground meaningful human-AI collaboration, rather than automation alone.

6.2 Background & Related Work

This section provides the theoretical and empirical foundation for our study. We begin by justifying our selection of the Flower and Hayes cognitive process model as our analytical lens. We then review the evolution of AI-assisted writing systems, followed by a discussion of how these systems affect writers' sense of agency and ownership over the writing process.

6.2.1 Theories on Writing Processes

Early conceptualizations of writing processes by Rohman [292] focused on the temporal evolution of written documents. Rohman introduced a three-stage model emphasizing "pre-writing" – the preparatory phase where writers engage in thinking and analysis to discover patterns in their subject matter. This stage, followed by "writing" and "re-writing," was seen as essential for producing what Rohman termed "good writing" (i.e., text that makes original and insightful contributions). While groundbreaking, this linear approach would later be challenged by more dynamic models.

Flower and Hayes [94] reconceptualized writing as a set of cognitive processes that writers deploy dynamically rather than in temporal stages. Their model identifies four primary processes: *planning* (i.e., constructing internal representations of knowledge through generating ideas, organizing ideas, and goal-setting), *translating* (i.e., transforming structured information into linear prose), *reviewing* (i.e., evaluating and revising text according to established goals), and *monitoring* (i.e., overseeing, regulating, and coordinating the writer's cognitive activities, such as deciding when to shift between planning, translating, and reviewing, and ensuring alignment with writing goals). These processes operate within a task environment that includes the rhetorical problem and the emerging text, drawing upon the writer's long-term memory for topic knowledge and audience awareness. The

processes form a hierarchical network where writers can move between processes at any time, or between high-level and local operational goals.

Nystrand [244] expanded the theoretical landscape by incorporating social dimensions into writing process analysis. His framework emphasizes writing as a communicative event where meaning is actively constructed between writer and reader within discursive communities. Nystrand argued that skilled writers anticipate readers' expectations and manipulate their text to establish a temporarily-shared social reality. Hayes and Nash [127] detailed the cognitive architecture of planning, including planning by abstraction, analogy, and modeling. Kellogg [158] explored the role of working memory in writing processes, while Hayes [126] expanded the original Flower and Hayes model to encompass social and physical environments, affect, and motivation. These contributions added depth to specific aspects of the writing process while building upon earlier foundational frameworks.

Our analysis employs the Flower and Hayes (1981) model as our primary theoretical lens for several reasons. First, it provides a comprehensive process model that describes writing behaviours. Second, its processes provide an analytical framework sufficient for examining the collaborative writing process between humans and AI systems. While Nystrand's model analyzes the social relationship between writer and reader, our research focuses on the collaborative interactions during writing, i.e. how humans and AI jointly engage in planning, translating, reviewing, and monitoring processes. The Flower and Hayes model allows us to examine how these cognitive processes are distributed and negotiated between human writers and AI systems during composition. Compared to other writing process theories, the Flower and Hayes model maintains an optimal balance between sophistication and analytical utility for our specific research context.

6.2.2 AI-Assisted Writing

Research on AI-assisted writing systems traces back to early implementations focused on creative writing support. Pre-transformers [352] systems like Creative Help [291] and Say Anything [334] utilized case-based reasoning and story repositories to generate context-aware sentence suggestions. Clark's [59] work examining user experiences with AI writing prototypes revealed that while participants found AI collaboration satisfying, the resulting text quality did not surpass that of unaided human writers. These early systems laid the groundwork for understanding both the potential and limitations of AI writing assistance.

The emergence of transformer-based large language models in 2017 catalyzed research into AI writing assistance. In creative writing, researchers have developed systems supporting story writing [379], playwriting [225], and character development [272, 297]. New systems support higher-level writing tasks such as prewriting [358], and generating perspective-specific feedback [22]. Specialized creative applications have emerged for tasks including metaphor generation [161], collaborative storytelling [240], and personal diary writing [166] as well as auxiliary creative tasks such as caption generation [156], title creation [252], and writing reflective summaries [70]. Technical writing applications have focused on enhancing accessibility and supporting specialized writing tasks, including peer review [355], literature reviews [55], and writing support for users with dyslexia or stuttering [101, 111].

User studies reveal complex dynamics in how writers interact with and perceive AI writing assistance. At a system interaction level, the design of AI suggestions significantly impacts user behaviour and output: sentence-level suggestions promote original content creation, while paragraph-

level suggestions improve efficiency [97]. Writers' engagement with AI assistance is also influenced by their personal values and goals. Writers show varying receptivity to AI support based on their confidence levels, demonstrating higher acceptance in areas where they lack expertise [25], and their desires for support are closely tied to their perception of support actors and personal values [100]. Moreover, this human-AI writing relationship raises important concerns. Studies by Jakesch et al. [145] and Poddar et al. [270] reveal that biased AI models can influence not only the resulting text but also users' own opinions. While users often value AI writing assistance highly, particularly for creative tasks [202], professional writers note persistent challenges with AI systems' ability to maintain consistent style and voice [141]. These findings highlight a central tension: as AI writing systems become more sophisticated, they must balance providing assistance while preserving authenticity and agency.

6.2.3 Agency and Ownership in AI-Assisted Writing

Recent research has examined how AI writing assistance affects users' sense of agency (i.e., their perception of control and autonomy over the writing process) and ownership (i.e., their feeling of personal investment in and attribution of the final text). Studies have shown that writers' sense of agency is significantly impacted by the level and type of AI intervention in the writing process. Robertson et al. [288] found that autocomplete suggestions could threaten users' autonomy. Similarly, Dhillon et al.'s [75] research demonstrated that while next-paragraph suggestions improved writing quality, longer AI text completions decreased satisfaction by undermining writers' independence. This finding aligns with Draxler et al.'s work [80], which showed that increased AI support corresponded with decreases in users' perceived control.

The relationship between AI assistance and text ownership is influenced by multiple factors, particularly professional context and writing purpose. Lee et al. [192] identified a direct correlation between self-reported ownership and the proportion of user-written versus AI-generated text. Biermann et al. [25] found that storywriters who emphasized the expressive and emotional value of writing insisted on maintaining direct control over translation, viewing this control as essential to preserving their writerly identity and integrity. Gero et al.'s research [100] revealed that the idea generation phase can particularly threaten ownership, with some writers considering the struggle with writer's block as integral to their writerly identity.

Several studies have identified factors that influence users' sense of agency and ownership in AI writing systems. Kobiella et al [174] found that participants who viewed AI as an enhancement tool rather than a replacement reported stronger feelings of accomplishment and ownership, while those who perceived their contributions as minimal experienced diminished ownership. Rezwana et al.'s [285] work highlighted that ownership perceptions depend on both contribution levels and leadership in the writing process, suggesting that interaction designs that maximize user agency can enhance ownership. These findings indicate that maintaining user agency and ownership requires careful consideration of interaction design, user control mechanisms, and the balance between AI support and user autonomy.

These investigations have demonstrated the need for continued focus on users' senses of agency and ownership when writing with AI. However, there are currently no broad reviews of the AI-assisted writing research landscape that have evaluated HCI researchers' and system designers' strategies against users' needs for preserving their agency and ownership throughout their writing

process.

6.3 Study 1: Reviewing Writing Process Dimensions in the Literature

To answer **RQ1:** "What design strategies are used or suggested in existing AI-assisted writing research, particularly in terms of interaction models and the intended use of AI outputs, and how are these strategies distributed across writing processes and writing contexts?", we conducted a PRISMA systematic literature review on the ACM Digital Library database, and coded the resulting paper dataset, guided by the Flower & Hayes Cognitive Process Theory of Writing [94] and writing contexts, interfaces, and interactions enumerated by Lee et al. [193]. We then performed thematic analysis [34] on the coded dataset in order to identify design strategies characterized by different interaction models, levels of AI support, and treatment of AI outputs. Finally, we coded the systems in our dataset by strategy in order to determine the distribution of the strategies across writing processes and contexts.

6.3.1 Methods

We conducted a systematic literature review following PRISMA guidelines [253] to identify and analyze research on AI writing support systems. PRISMA (**P**referred **R**eporting **I**tems for **S**ystematic **R**eviews and **M**eta-**A**nalyses) is a series of systematic review guidelines that are intended to improve the reporting and replicability of scientific literature reviews and meta-analyses. Our review focused on papers published between 2018-2024, corresponding to the emergence and widespread adoption of transformer-based language models [352]. This period is marked by the explosion of AI research in HCI, visible in the publishing dates of papers in our dataset Figure 6.2a.

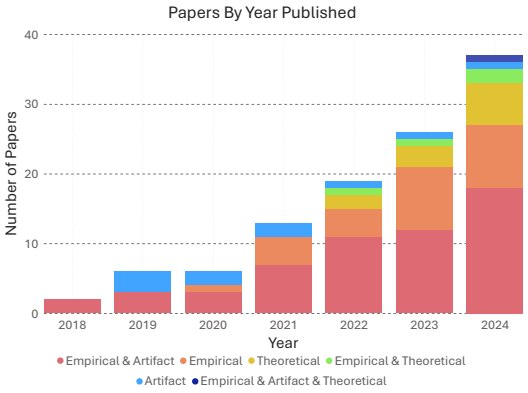
Query Construction

We developed our search query through an iterative process, beginning with a set of seed papers and expert knowledge in the field. We expanded our initial keyword list through multiple refinement cycles. The final search query combined writing-related terms with AI-related terms in order to capture as many potentially-relevant papers as possible:

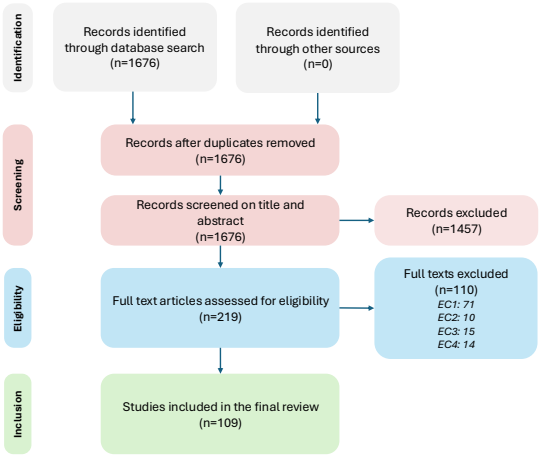
```
("writing" OR "writer" OR "write" OR "collaborative" OR "collaboration"
OR "collaborate" OR "collaborating" OR "author" OR "authors" OR
"creativity support" OR "co-creation" OR "co-writing") AND
("AI" OR "language model" OR "artificial intelligence" OR "generative"
OR "chatbot" OR "natural language processing" OR "NLP" OR "LLM" OR
"digital assistant")
```

Exclusion Criteria

We limited our review to peer-reviewed papers published in English, including journal papers, conference proceedings, and extended abstracts. We developed four primary exclusion criteria:



(a) Distribution of papers by year, color-coded by HCI contribution type, highlighting the rapid growth of AI-assisted writing research following the introduction of transformer-based language models.



(b) PRISMA flowchart illustrating the paper selection process for Study 1, including identification, screening, eligibility, and inclusion of 109 papers from an initial set of 1,676 records.

Figure 6.2: Study overview: distribution of selected papers and paper selection process.

EC1. Papers where AI interaction is not providing AI-assisted writing support, defined as AI writing with a user to create a natural language written artifact.

EC2. Papers presenting purely technical, backend, or algorithmic contributions without user interaction.

EC3. Papers focusing on non-natural language output formats (i.e., code or images exclusively).

EC4. Papers that did not present a user study, artifact or system contribution, theory or conceptual framework, or systematic review.

Database Selection

To determine the optimal database for our review, we conducted a preliminary analysis across multiple digital libraries. We systematically sampled 100 papers from each of ACM Digital Library, IEEE Xplore, Taylor & Francis, and Wiley by using a random sampling to select papers from search results for our query with 2018-2024 publication dates. We then applied our exclusion criteria to these papers' titles and abstracts. This preliminary assessment was designed to evaluate the concentration of relevant literature across databases. The ACM Digital Library yielded significantly more relevant results (11%) compared to IEEE (2%), Taylor & Francis (3%), and Wiley (3%). Given this substantially higher concentration of relevant publications, we determined that the ACM Digital Library would provide the most comprehensive and targeted corpus of literature addressing our research questions.

Screening Process

Our initial search yielded 1,676 papers. One researcher conducted the initial screening, applying our exclusion criteria to titles and abstracts, which identified 219 papers for full-text review. To ensure

reliable coding, we conducted an inter-rater reliability test on a random sample of 25 papers from this set. Two researchers independently coded these papers based on the full text, achieving a Cohen's kappa of 0.84, indicating strong agreement[222]. We resolved disagreements through discussion and consensus with a third researcher.

Following the confirmation of inter-rater reliability, we divided the remaining papers between two researchers for independent full-text review. This process resulted in the exclusion of 110 papers: 71 for not providing co-writing support (criterion 1), 10 for purely technical contributions (criterion 2), 15 for non-natural language output (criterion 3), and 14 for not meeting our paper type criteria (criterion 4). Our final dataset comprised 109 papers. Figure 6.2b provides an overview of the paper selection process.

Analysis

We employed a codebook thematic analysis approach, developing our initial codes from Flower & Hayes' cognitive process model [94] and Lee et al.'s design space framework [193]. Following King & Brooks [169] and Braun & Clarke [34] we established our coding framework early in the process allowing us to inductively identify rich qualitative themes from the data.

6.3.2 Study 1 Findings

We first describe how writing support in the literature varies across five writing contexts, highlighting differences in the goals, users, and processes emphasized in each. We then introduce four overarching design strategies that characterize how systems support writers across these contexts, each with distinct implications for agency, task delegation, and interaction design.

Writing Context Characteristics

In the following section we characterize our dataset by the writing context where they offered support, based on contexts adapted from Lee et al. [193]. We describe each writing context and how AI-assisted writing research supports each cognitive writing process across them; we also report how many papers[1] were coded into each writing context, as shown in Table 6.1.

1. **Academic (31 Papers).** The Academic writing context includes papers that are focused on research, analysis, or educational use. Papers in this context include topics such as assistance with literature review [55, 359], peer review [236, 333], academic writing [336, 321, 39, 311], and essay writing [70, 357, 3, 281, 22]. AI-assisted writing systems in this context are often focused on structured skill development for the users. Support for planning processes are typically intended to help users connect and structure their own ideas to accomplish complex tasks [281, 336]. Translating support provides preliminary drafts or helps the user to restructure their ideas in a different form (e.g. point form to prose), but encourages the user to write and integrate ideas on their own [333, 321, 56]. Reviewing support is delivered in qualitative form, such as suggestions or summaries [308, 22, 321]. Systems do not revise the user's text directly, instead recommending improvements to prompt the user to revise the work themselves.

[1]Note that the counts of papers do not add up to 109, the number of papers in our dataset. Although most papers only had a single context, a small number spanned multiple contexts.

Cognitive Processes		Writing Contexts				
		Academic	**Creative**	**Formal**	**Personal**	**General**
Planning	*Generating*	[236, 321, 55, 39, 311, 302]	[252, 182, 192, 59, 297, 98, 266, 275, 362, 57, 379, 225, 240, 304, 156, 272, 102, 71, 25, 139, 100, 322, 342, 29, 202, 358, 167, 138]	[161, 99, 8, 97, 24, 78, 113]	[8, 192, 312, 362, 385, 61, 75, 43, 150, 166, 202]	[50, 98, 195, 40, 76, 61, 24, 332, 282]
	Organizing	[70, 3, 281, 336, 333, 55, 311, 302, 359]	[182, 297, 362, 57, 379, 376, 102, 165, 25, 144, 322, 358]	[161, 165, 78, 144]	[356, 312, 362, 385, 204, 166, 145, 270]	[111, 76, 332, 282, 204]
	Goal-setting	[357, 281, 336, 22, 55]	[182, 272, 377, 25, 100, 342]	[215]	[385, 43, 150, 22]	
Translating		[321, 333, 56, 55, 39, 105, 198, 263, 311, 77]	[252, 266, 57, 102, 71, 377, 25, 139, 100, 322, 29, 202, 167, 138]	[161, 99, 97, 24, 119, 208]	[385, 204, 43, 150, 166, 145, 208, 81, 202]	[24, 332, 204]
Reviewing	*Evaluating*	[355, 308, 296, 357, 143, 268, 336, 22, 321, 55, 5, 198, 311, 319, 278, 359]	[266, 362, 134, 376, 377, 100, 203, 29, 373, 52]	[243, 288, 238]	[355, 356, 265, 362, 162, 22, 288, 52]	[58]
	Revising	[355, 308, 143, 268, 118, 22, 55, 136, 39, 47, 72]	[379, 134, 29]		[355, 371, 265, 385, 204, 22]	[101, 111, 76, 282, 204]
Monitoring		[357, 281]	[203]	[8, 295]	[8]	[58, 232]

Table 6.1: Mapping of cited papers to writing processes and writing contexts, based on our systematic review of AI-assisted writing literature. The table reveals uneven research attention across cognitive processes and contexts–for example, strong representation of Generating and Translating activities in Creative and Academic settings, and limited focus on Monitoring across all contexts.

2. **Creative (37 Papers)**. Creative writing papers focus on artistic expressions and narrative-based texts. In the Creative writing context, topics include: story writing [252, 192, 59, 57, 379, 102, 165, 25, 322], collaborative storytelling with AI [182, 240, 377], including CSCW work on AI support for human collaborative storytelling [304] and using dialects in creative writing [364]. Other assistance includes character creation [297, 272], poetry [29], lyric generation [275], writing screenplays [225], and design fiction [342]. AI also provides support with rhetorical or stylistic elements such as forming metaphors [98] or learning vocabulary [266]. A focus of researchers in this area is conducting empirical studies with writers to discover their writing strategies and requirements for support [25, 139, 100, 202, 358, 167]. Support for planning takes the form of generative ideation, usually presented as suggestions [362, 192], although some systems have a more equal and collaborative storytelling focus that weaves the AI ideas into the story text [252, 182]. Support for translating often occurs simultaneously with support for generating ideas, creating narratives or creative elements that blend the user's prior text with new ideas from the AI [102, 57]. Support for reviewing features a mix of quantitative and qualitative feedback, with a focus on the AI evaluating text and providing suggestions rather than revising it directly [336, 55].

3. **Formal (16 Papers)**.The Formal writing context represents professional, standardized modes of writing, characterized by structured forms, limited use of personal or emotional expression, and purpose-driven tasks that entail specific communicative goals. AI support from papers in this context is focused on topics like enhancing productivity [8, 215], writing business emails or reports [97, 208, 238], reviews [24], professional design problems [78], copywriting [165], document analysis [144], clinical use [119] and creating solutions to business problems [113]. Planning support in this context is focused on extending and organizing the user's ideas, often through analogies and cross-domain reasoning [161, 99]. Translating support is focused on writing efficiency, enabling the AI to write in the same interaction location as the user, or to make suggestions that are integrated directly in the text [97, 24]. Finally, reviewing support was limited, and focused on evaluation using quantitative feedback like readability metrics [238], and visual feedback such as progress bars [243].

4. **Personal (24 Papers)**. The Personal writing context concerns self-expression and sharing one's thoughts, feelings, and experiences. Compared to other contexts, it embraces informality, subjectivity, and authenticity. Writing tasks in this context include non-academic opinion essay writing [8, 356, 192, 385, 75, 202], blog or social media posts [312, 61, 204, 43, 150, 22, 145, 270], personal messages [168] and journalling [166]. AI support in this context is generally targeted at lay users, emphasizing ease-of-use. Planning and translating support are intermingled due to the frequent use of longer AI outputs that directly ideate and write for the user, though these are generally presented as suggestions in order to preserve the user's engagement with the text [43, 166]. We also see transformation of user inputs between modalities like speech to text or visuals to text [385, 204], or between textual forms like keywords to prose [166]. Reviewing support is typically provided through quantitative feedback, with a focus on evaluation rather than direct revision [355, 265].

5. **General (15 Papers)**. The General writing context contains systems that are presented for use in multiple contexts, or where the system design is not adapted to solving problems from

a particular contextual domain. We also included systems that provide accessibility support in this context. Writing tasks include dyslexia support [111], support for people with speech impediments [101], writing both personal and professional emails [50, 195, 40], and writing applications which are targeted at multiple contexts [76, 24, 282, 204, 332]. A recurrent theme in planning and translating support in this context was the provision of interfaces that enabled rapid iteration and organization of idea and text generations [204, 282, 332]. Systems commonly provided suggestions for revisions which could be integrated directly into the writing area [50, 195, 40], aiding efficiency and idea exploration in the text.

Strategies for AI-Assisted Writing Support in HCI Research

Our analysis revealed four overarching design strategies for AI writing support that span cognitive processes and writing contexts. These strategies are distinguished by the AI's role, intended user behaviors, interaction outcomes, interface design, and usage of AI outputs. Systems can combine elements from multiple strategies based on their supported writing processes and contextual requirements. Each strategy offers varying support for writers' sense of agency, ownership, and task delegation preferences across different contexts.

1. **S1: Structured Guidance**. This strategy represents a scaffolding approach where AI systems function as writing coaches or tutors, guiding users through document development while maintaining their autonomy and preserving agency. This strategy emphasizes active skill development through structured practice rather than passive reception of AI-generated content, typically requiring predefined writing tasks. The strategy comprises four key components. **Pattern Mapping** focuses on developing connections and pattern recognition within existing content rather than generating new ideas, with AI systems helping users locate patterns in their data and analyze potential suggestions. **Sequential Development** denotes an iterative approach through drafts and milestones, where the system guides users in adapting suggestions to build their writing capacity. **Scaffolded Feedback** delivers assessments through structured templates, combining quantitative metrics with clear evaluation frameworks, and encouraging the user to perform their own revisions. Finally, **Workspace Control** employs user interfaces that physically separate AI and user workspaces, ensuring users maintain control over textual changes while explicitly initiating support requests at each stage of the writing process. Revisions utilizes proposals which the user can reference, or analysis to help the user revise their text, which ensures the user still contributes to the text. This approach respects writers' need to maintain agency over ideation and organizing, thereby preserving their sense of ownership.

2. **S2: Guided Exploration**. This strategy positions AI systems as facilitators that enable users to actively explore and make connections within an idea space, with the AI functioning as both map-maker and guide. This strategy supports both well-defined and ill-defined writing tasks, emphasizing user engagement through iterative exploration and selection while maintaining creative control. It encompasses four main components. **Idea Navigation** implements a structured, self-directed approach that balances assistance with skill development, focusing particularly on interfaces which allow users to swap between generations to explore different approaches to their rhetorical problem. These systems enumerate the idea space using ideas generated by the AI. **Output Variation** denotes the provision of multiple types of

output by the AI (i.e., narrative elements like plot and creative elements like dialogue), offering flexibility in AI generation. Systems directly replace user text in the writing area, enabling users to evaluate revisions in place, with the option of using the exploration interface to undo changes. **Iterative Revision** utilizes the map of the idea space generated in exploration to both structure potential ideas and guide revision, facilitating an iterative model of exploration and refinement by the user. **Proposal Integration** maintains user control by having the system present ideas and text generated by the AI as proposals. The focus on exploration offers the user flexibility in how to integrate generations into the artifact, with an emphasis on user-initiated AI output. This balances task delegation needs by allowing writers to maintain agency over idea selection while delegating generation, supporting their sense of ownership.

3. **S3: Active Co-Writing**. This strategy establishes AI systems as active writing partners, enabling a collaborative relationship where users selectively offload writing tasks while maintaining editorial control over the final output, though with potential implications for ownership. This strategy accommodates both well-defined and ill-defined tasks by supporting rapid iteration and efficient workflows. It consists of five primary components. **Direct Generation** involves direct generation of substantial content (i.e., full drafts or long text completions) intended for integration into the final artifact, encompassing both idea development and formal aspects of the text. **Content Conversion** preserves user ideas through various transformation types (i.e., foreign language translation, translating keywords to prose). The transformation retains the user's original meaning, utilizing the AI to deliver that meaning in new forms. **Efficiency Optimization** denotes prioritization of speed and usability through streamlined interactions, contrasting with skill-development approaches. These are often deployed in Professional contexts where productivity is paramount. **Turn-based Creation** denotes turn-based interactions through chat or collaborative storytelling, facilitating human and AI creative input with automatic integration of AI contributions into the final artifact. Finally, **Result Ownership** maintains user control through suggestion selection and user-initiated AI output. However, unlike Proposal Integration, suggestions are integrated directly in the final text which may challenge writers' sense of agency by blurring task delegation boundaries.

4. **S4: Critical Feedback**. This strategy positions AI systems as editors and organizers, facilitating a user's reflective practice through structured feedback while maintaining a deliberate separation between the creation and analysis phases, supporting clear task delegation boundaries. This strategy requires well-defined tasks to enable evaluation and comprises four components. Unlike strategies that span the entire writing process, Critical Feedback represents a specialized approach where systems focus on reviewing and evaluation, maximizing analytical depth through structured assessments. **Qualitative Feedback** implements anthropomorphized or less-structured interactions that simulate tutoring scenarios through chat or natural language feedback. This method can provide revisions, but typically requires manual integration of suggestions by users. **Quantitative Analysis** provides structured assessments with a stronger focus on evaluation than revision, utilizing numerical or visual feedback. **Hybrid Evaluation** combines qualitative and quantitative approaches, using formal templates rather than conversational formats. This method offers a balance of revision and evaluation

support that protects users' agency by requiring effort to integrate into the text. **Revision Guidance** connects analysis and organization by offering revision suggestions based on idea summaries and providing fine-grained tools for specific revision tasks (i.e., merging, rewriting, summarizing). **Analysis Separation** maintains user control through deliberate separation between AI output and user workspace, requiring user-initiation of AI output, and introducing friction by requiring manual integration of AI-proposed revisions. This design choice deliberately preserves the writer's agency over implementation decisions, reinforcing their ownership of the final text through strategic task delegation.

We applied our strategy framework to code the papers presenting system contributions (n=62) from our dataset to characterize the landscape of existing research systems. A single researcher assessed each system against the defining characteristics of each strategy. This coding revealed that S3 (Active Co-Writing) was the most commonly deployed approach (23 systems, 37.1%), followed by S1 (Structured Guidance) (19 systems, 30.6%), S2 (Guided Exploration) (11 systems, 17.7%), and S4 (Critical Feedback) (9 systems, 14.5%), with the full distribution shown in Table 6.2. Clear patterns emerged across writing processes, with S1 dominating Evaluating (54.5%) and S3 leading in Generating (52.9%) and Translating (50%) processes. Context-specific preferences were also evident, with Academic writing favoring S1 (61.1%), Creative writing employing S3 (50%), and Formal writing preferring S2 (42.9%). Creative writing showed surprisingly high deployment of S3 systems, which are the most likely to threaten ownership. While the single-coder approach represents a limitation, this application of our framework highlights opportunities for more nuanced strategy implementation across cognitive processes.

6.4 Study 2: Investigating AI's Influence on Ownership in Writing

To answer **RQ2**: "Which cognitive processes do writers consider essential to control in order to maintain their sense of agency during AI-assisted writing, and how do user situations, writing contexts, and AI interaction types shape their perceptions of ownership?", we conducted interviews with 15 writers.

6.4.1 Methods

We detail the methodological details of our second study, including participant recruitment, study procedures, and our approach to data analysis.

Participants

We recruited 15 writers (8 women, 6 men, 1 did not specify; other gender options were offered) across two age groups: 5 participants aged 18–24 and 10 aged 25–34. Participants were based in North American and European cities and were recruited via social media and email invitations. They possessed diverse writing experience, including academic research papers (W11, W14), knowledge translation (W4), short stories (W2), poetry (W9), novels (W7, W13), essays (W1, W10), blogs (W4), screenplays for TV shows (W12), newspaper articles (W15), personal diaries (W5), internal

S1: Structured Guidance S2: Guided Exploration S3: Active Co-Writing S4: Critical Feedback

Cognitive Processes		Writing Contexts					Total
		Academic	Creative	Formal	Personal	General	
Planning	*Generating*	S1: 1 system (33%) S3: 2 systems (67%)	S1: 2 systems (13%) S2: 5 systems (31%) S3: 8 systems (50%) S4: 1 system (6%)	S2: 2 systems (67%) S3: 1 system (33%)	S1: 2 systems (25%) S3: 5 systems (63%) S4: 1 system (13%)	S2: 4 systems (67%) S3: 2 systems (33%)	S1: 5 systems (15%) S2: 10 systems (29%) S3: 18 systems (53%) S4: 1 system (3%)
	Organizing	S1: 5 systems (83%) S4: 1 system (17%)	S1: 1 system (13%) S2: 2 systems (25%) S3: 4 systems (50%) S4: 1 system (13%)	S2: 1 system (100%)	S1: 3 systems (50%) S3: 2 systems (33%) S4: 1 system (17%)	S2: 3 systems (60%) S3: 2 systems (40%)	S1: 9 systems (38%) S2: 6 systems (25%) S3: 7 systems (29%) S4: 2 systems (8%)
	Goal-Setting	S1: 4 systems (80%) S4: 1 system (20%)	S2: 1 system (25%) S3: 3 systems (75%)	S2: 1 system (100%)	S1: 1 system (25%) S3: 2 systems (50%) S4: 1 system (25%)	No systems	S1: 5 systems (38%) S2: 2 systems (15%) S3: 5 systems (38%) S4: 1 system (8%)
Translating		S1: 2 systems (50%) S3: 2 systems (50%)	S1: 1 system (20%) S2: 2 systems (40%) S3: 2 systems (40%)	S2: 2 systems (67%) S3: 1 system (33%)	S1: 1 system (20%) S3: 4 systems (80%)	S2: 1 system (50%) S3: 1 system (50%)	S1: 4 systems (22%) S2: 5 systems (28%) S3: 9 systems (50%)
Reviewing	*Evaluating*	S1: 7 systems (64%) S3: 2 systems (18%) S4: 2 systems (18%)	S1: 2 systems (33%) S3: 2 systems (33%) S4: 2 systems (33%)	S1: 1 system (50%) S4: 1 system (50%)	S1: 3 systems (50%) S4: 3 systems (50%)	No systems	S1: 12 systems (55%) S3: 4 systems (18%) S4: 6 systems (27%)
	Revising	S1: 5 systems (71%) S3: 1 system (14%) S4: 1 system (14%)	S3: 1 system (50%) S4: 1 system (50%)	No systems	S1: 2 systems (33%) S3: 1 system (17%) S4: 3 systems (50%)	S2: 2 systems (40%) S3: 2 systems (40%) S4: 1 system (20%)	S1: 6 systems (35%) S2: 2 systems (12%) S3: 4 systems (24%) S4: 5 systems (29%)
Monitoring		S1: 2 systems (100%)	No systems	S3: 1 system (100%)	S3: 1 system (100%)	No systems	S1: 2 systems (67%) S3: 1 system (33%)
Total		S1: 11 systems (61%) S3: 4 systems (22%) S4: 3 systems (17%)	S1: 3 systems (15%) S2: 5 systems (25%) S3: 10 systems (50%) S4: 2 systems (10%)	S1: 1 system (14%) S2: 3 systems (43%) S3: 2 systems (29%) S4: 1 system (14%)	S1: 5 systems (33%) S3: 6 systems (40%) S4: 4 systems (27%)	S2: 4 systems (44%) S3: 4 systems (44%) S4: 1 system (11%)	S1: 19 systems (31%) S2: 11 systems (18%) S3: 23 systems (37%) S4: 9 systems (15%)

Table 6.2: Distribution of systems by strategy across writing processes and contexts to show the prevalence of each design strategy in the literature dataset. Cell colouring is proportional to the prevalence of strategies deployed for systems in that cell, subject to a minimum height for readability. Systems could be coded to more than one process or context.

project documentation (W6) and creative fiction (W3, W8). 11 participants reported having professional writing experience (i.e., when writing is paid or a core part of their occupation). Weekly time spent on writing varied, with 5 participants writing 1–4 hours, 5 writing 4–7 hours, 3 writing 7–10 hours, and 2 spending more than 15 hours per week.

Given our focus on AI-assisted writing, prior experience with AI writing tools was an inclusion criterion. All participants reported using ChatGPT, with Grammarly and Microsoft Copilot being the next most popular tools. Some advanced users also experimented with other LLMs and specialized AI writing tools, including Claude, LLaMA , and writing tools like Sudowrite. To ensure our findings were not biased toward users with a particular level of knowledge of generative AI, we recruited writers with varying generative AI expertise, ranging from slightly to extremely knowledgeable.

Procedure

Each study session lasted between 60 and 90 minutes and was conducted online via recorded video calls by the lead author, allowing us to reach participants across multiple geographic locations. We introduced participants to the study and then asked them to complete a 5-minute pre-survey to provide consent and share demographic information, their writing experience, and AI usage. We informed participants of their right to withdraw from the study at any time and compensated each participant with 20 CAD for their time. The institution's research ethics board approved the study protocol.

Following the pre-survey, we conducted a semi-structured interview in which participants described their writing background and experience with AI, and shared their perspectives on how AI influences their sense of ownership across each aspects of the writing process. We defined each process, to ensure writers could relate their practices to the processes. Finally, participants completed a 10 minute post-interview survey, reflecting on the discussion and rating 16 Likert-scale statements (4 for each process). This survey helped us gauge preferences for AI involvement in each element of the writing process. Further details on the post-interview survey Likert items are provided in Figure 6.3).

Data Analysis

The data included transcripts of the interview recordings and responses to pre- and post-interview surveys. To identify factors that influence writers' sense of ownership in the AI-assisted writing process, we conducted a reflexive thematic analysis [34] of transcripts through an inductive-deductive approach. Guided by the cognitive process theory of writing, we used the main writing processes—planning, translation, reviewing, and monitoring—as predefined codes to structure our interpretation, while also inductively identifying new patterns. The pre-survey data provided important context about each participant's background in writing and prior experience with AI tools. The post-interview survey helped quantify attitudes toward AI across different cognitive processes. Scores for negatively worded items (Q1, Q3, Q5, Q9, and Q13) were reversed (see Figure 6.4). Given the varied perspectives on ownership across writing elements, our goal was not to aggregate results into a single measure of ownership and agency but rather to examine distinct aspects of the writing process.

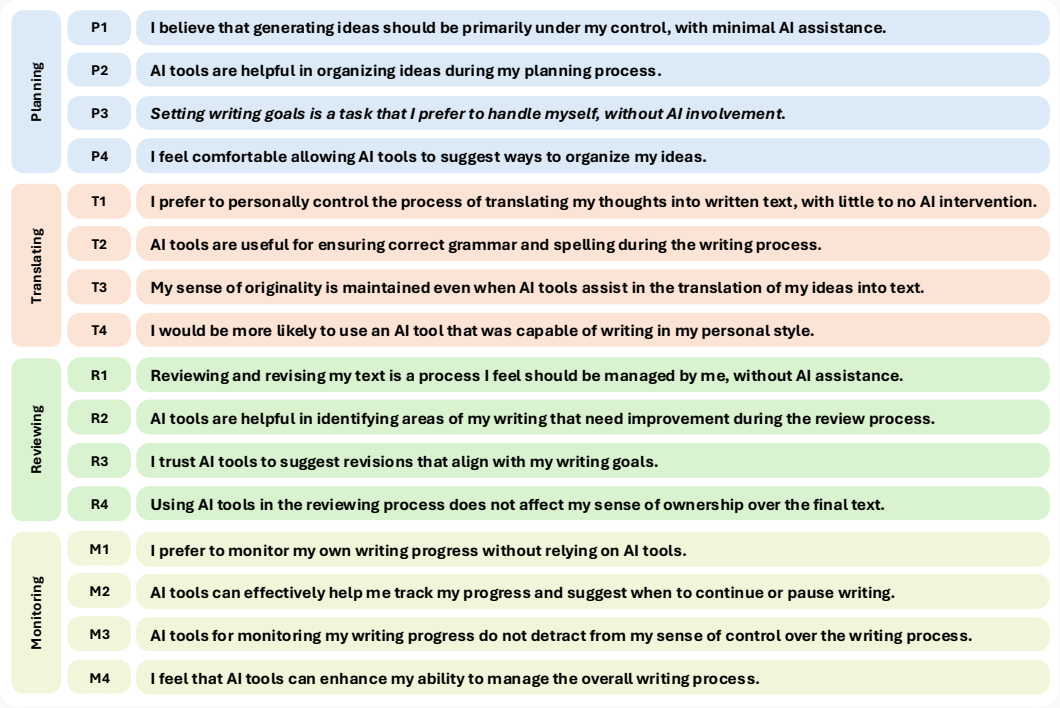


Figure 6.3: Likert-Scale Statements on User Perceptions of Cognitive Processes during Writing

6.4.2 Study 2 Findings

The post-interview survey responses are summarized in Figure 6.4. Items are grouped into sets of four, labeled **P**1-4, **T**1-4, **R**1-4, and **M**1-4, corresponding to the four cognitive processes in the Flower and Hayes writing model: **P**lanning, **T**ranslation, **R**eviewing, and **M**onitoring. The item questions are detailed in Figure 6.3. The distribution of ratings reflects a range of perspectives on the extent to which writers want AI to intervene across different processes. We interpret this diversity through our thematic analysis, and share findings on how writers perceive and maintain a sense of ownership and agency over the writing process when working with AI. We group these insights under three primary themes, each highlighting a different dimension of the writers' relationship to AI and their work:

1. **When Ownership Matters**: This theme delineates the contextual factors–such as time constraints, level of trust in AI, task importance, and perceived competence–that shape writers' decisions around how much control they want to retain and how much they are willing to delegate to an AI tool, even if it means their sense of ownership is encroached. Instead of assuming the desirability of ownership as an inherent or static prerequisite, this theme showcases the flexible role that human agency plays in AI-assisted writing and how it responds to situational factors. It also highlights situations where the risk of writers' overreliance on AI is particularly prevalent.

2. **What Writers Want to Own**: This theme characterizes the aspects of the composition process and product from which writers derive their sense of ownership and prioritize as their

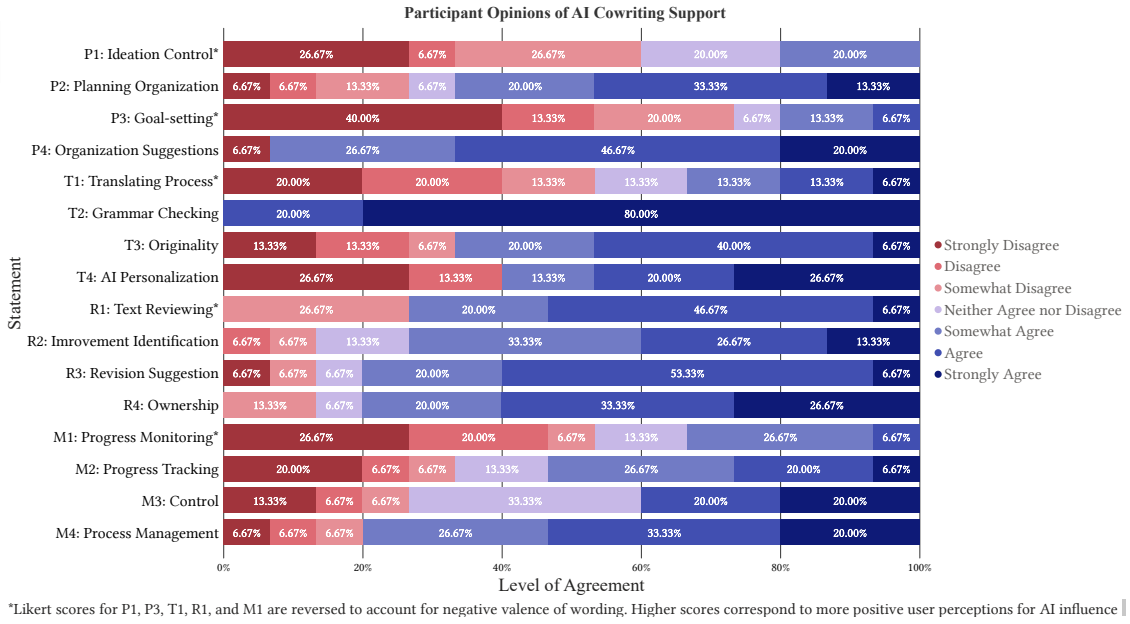


Figure 6.4: User perception likert-Scale items on writers' sense of ownership across cognitive processes [94]. The distribution shows notable variation in the desirability of AI support across processes.

primary contribution. We identify a central distinction between content and form: writers prioritize idea generation and planning as their primary contribution in content-oriented writing, where the purpose is primarily expository, while in form-oriented writing, where the focus is on style and voice, they emphasize the need to exercise more control during translation and revision to convey their unique expression.

3. **How AI Interactions Shape Ownership**: This theme explores how interaction design impacts writers' senses of agency and ownership. We look at how different interface elements shape how writers feel when AI intervenes, such as the option to receive suggestions rather than direct edits, providing multiple suggestions, exercising final say, and UI affordances for enabling and disabling AI input. This theme highlights the critical role that Human-AI interaction design can play in maintaining writers' sense of agency and ownership in AI-assisted workflows.

Together, these three overarching themes offer a way to grapple with the complex interplay between writer agency, task demands, and AI functionality, helping AI system designers make sense of how ownership is negotiated and maintained in AI-assisted writing.

When Ownership Matters

We found four factors that influence how much control writers are willing to give to the AI and the extent to which they care about maintaining their sense of ownership in the first place – **time** constraints, task **importance**, **confidence** in the writers' own abilities, and **trust** in the AI's capabilities.

1. **Time:** A key reason writers are drawn to generative AI tools is efficiency. Therefore, in time-sensitive situations, writers are more willing to delegate tasks to the AI. W4 described

ChatGPT as *"a huge time saver"*, noting how *"it sometimes helps when you're working on something super last minute, to have an AI look at it as well, and go through it in greater detail and precision"* than them. In addition to proofreading, writers are also more willing to delegate other aspects of the writing process. W11 shared how they used AI tools to transform rough bullets into polished writing. *"There are also situations where I'm running short of time, and I will have a list of things I want to add... ordered in a reasonable way as I want them to appear in the writing. Then I will just ask ChatGPT to draft something based on the list."* A similar point was echoed by W7, who described how they convert messy outlines into coherent text: *"to save time, I will write out all bullet points myself that are really messy, and then have ChatGPT turn it into a letter."*

2. **Importance:** Writers de-prioritize ownership in low-stakes tasks, such as routine emails or straightforward professional communication, where clarity and efficiency are the primary goals. Such tasks tend to be perceived by writers as more functional than creative, making AI tools more acceptable for generating content without affecting their sense of ownership. The inverse is also true–when the stakes are high, writers become much less open to the idea of AI involvement, as captured by W2's comment: *"It depends how important this project is, because if it's very, very important to me, I would give [AI] less responsibility, almost to the point where it's just used as like something that I accept or reject, just like an editor who works for you, you can either accept or reject it or revise it... if it was not that important, like an email that I'm just kind of sending off. I would give it almost all the work, honestly."* W15's remark reveals how this choice to adjust the importance of ownership is deliberate, and not necessarily due to a lack of self-awareness – *"I use [AI] to finish my emails when Gmail tells me to finish with yours sincerely... I'm like, sure that's what I say. So for me, that's the sort of use of LLMs that I find quite pervasive in the background, and which I am definitely happy to use... I suppose there's an irony in pressing Tab to write 'Yours sincerely'. You see, right? You're not being sincere."*

3. **Confidence:** For writers who feel less confident in specific language skills, AI tools can serve as a resource for checking for linguistic accuracy. Relying on AI for such help does not necessarily impact the writers' sense of ownership, as W3 observes: *"for grammar and spelling, those are inconsequential, right? I don't associate that with the voice... It's a menial task that can be taken care of by AI that doesn't impact someone's voice."* This selective reliance on AI allows writers to focus their energy and sense of ownership on other parts of the work, while using AI to polish weaker areas. W3 elaborates on this intentional boundary: *"I would want to make sure that anything that revolves around characters talking with one another, or whenever I write about the thoughts that the character is experiencing in their head, I'd want those to be my own work. But when it comes to describing a scene or a setting... that's something that as a writer, I'm not that great at, and so it's seeking help to make sure that my own work gets very polished."*

4. **Trust:** Just as confidence in their own abilities influenced writers' sense of ownership, their delegation choices were also shaped by their trust[2]—or lack thereof—in the AI's ability to deliver reliable output. This was particularly evident among writers who were confident in

[2]The concept of trust and agency are interrelated, as they both influence users' decision-making abilities [172].

their own abilities but skeptical of AI. For example, W10 explained, *"I don't particularly like the writing style. I don't trust it enough. There will always be a few nitpicks in any paragraph that I'll have with it. So in that way, I feel like I've been able to retain a total sense of ownership. I don't feel like it's influenced it any more than if I had someone read it and they said I liked it, or I didn't like it, or this part sucks."* The ability to critically evaluate AI output helped writers like W10 maintain a sense of separation between looking at AI output and feeling like it inadvertently influenced them. W8 held reservations about the AI's ability to gauge how humans would respond to a piece of writing – *"I just don't trust AI to judge whether something is understandable to a person or not, especially because of the variety of audiences I write for."* Instead, they turned to humans for feedback, relying on colleagues and friends with different levels of expertise, from both within and outside their fields, to get varied perspectives on their work.

Table 6.3: Delegation Strategies Based on Content and Form Contributions with Expanded Planning Categories

Cognitive Processes		**Content (e.g., Academic)**	**Form (e.g., Creative)**
Planning	*Generating*	Strong ownership over novel ideas with minimal AI input. W9: "The core part of writing is ideas... even if ChatGPT helps organize, the ideas remain mine."	AI-assisted ideation via prompts, but creativity retained by writer. W3: "I'd use [AI] for prompt creation... as a starter, then dive into writing."
	Organizing	AI supports in structuring ideas, retaining ownership over logic flow. W9: "I let ChatGPT organize the ideas... but the logical flow is my own."	AI-assisted outline creation for enhanced cohesion, writer's tone. W3: "Organizing can be AI-assisted if it doesn't alter my style."
	Goal-setting	Writer defines key objectives and frameworks; AI used for background structure alignment. W13: "If I outline the goals clearly, ChatGPT can help format, but the primary direction remains mine."	Creative goals set by writer; AI supports structure adjustments. W7: "My voice is in the goals of the story, AI only aids in structure refinement without altering intent."
Translating		AI used for drafting structured text; ownership tied to novel ideas, not genre standards. W7: "Academic writing feels less mine... I'm fine with delegating structure [to AI]."	AI assistance should be limited; primary voice retained through sentence-level decisions. W7: "Where I feel the most ownership is over sentences themselves."
Reviewing	*Evaluating*	Grammar and clarity editing delegated to AI for efficiency. W3: "For grammar, those are inconsequential... AI can handle it."	Limited AI assistance; primary voice retained through sentence-level decisions. W7: "Where I feel the most ownership is over sentences themselves."
	Revising	AI-intervention to refine writing clarity is welcome. W14: "I have a tendency to overwrite, explaining things in a longer way; it can be much more concise. [AI can help] in that stage of the editing process."	Strong sense of novelty in breaking conventions for stylistic effect and unique voice. W7: "I break grammatical conventions for aesthetic effect... it's important for the voice, so I'm not interested in AI changing those choices."

Legend: Red - Little Support Yellow - Moderate Support Green - Significant Support

What Writers Want to Own

In analyzing the areas from where writers draw their sense of ownership, we found a clear and recurring pattern: writers value ownership most strongly over components of the composition process they see as their primary contribution. Writers tend to be more open to delegating composition tasks

to an AI for areas that are tangential to their perceived primary contribution. When task delegation is done this way, writers' sense of autonomy and joy in creating something novel is maintained—or even enhanced in cases where the AI frees up their focus.

We identified two main types of contributions—content and form—each linked to specific cognitive processes. Content contributions involve generating ideas and setting goals, aligning with the planning process in the Flower and Hayes model [94]. Form contributions focus on style, tone, and flow, aligning with translation and revision. This content-form distinction connects writing contexts with cognitive processes: academic and non-fiction writers prioritized content to convey ideas clearly, while fiction writers emphasized form, valuing their unique voice and style. Below, we explore writers' sense of ownership for these two types (see Table 6.3 for mapping):

1. **Content:** When the purpose of a piece of writing is to convey pre-existing ideas or information with clarity–as is common in academic and non-fiction writing, the content itself becomes the primary contribution and the central focus of ownership for writers. In these writing contexts, writers derive a sense of ownership by engaging in cognitive processes involved in ideation and organization of ideas, making planning the dominant process that they seek to control. When the ideas are established, writers are open to using AI tools for translating ideas into clear language or reviewing their work to enhance clarity and polish. This is reflected in W9's perspective on non-fiction: *"I think the ideas are the core part of the writing. So if I'm giving ChatGPT the ideas that I want it to kind of organize, I think I still maintain that ownership of like, oh, those are my ideas, but it's enhancing my writing."*

   The desire to emphasize ownership over content rather than form is influenced by external constraints, such as word counts or stylistic conventions. W7 captures this sentiment when discussing the formulaic nature of research papers: *"I feel less ownership of my writing in general, just because the rhetorical context in which I'm writing is so rigid and has such clear expectations... I feel like I did something interesting as part of the research project, but the write-up itself is just a write-up and nothing more. [I'm] fine to delegate the vast majority of that process to someone else."* Since LLMs are trained on established standards, they can assist with refining text to meet these norms.

2. **Form:** In contrast, when writers have freedom over form, the idea of AI models intervening in the translation process becomes less appealing. W7, a professional literary fiction writer and novelist, explained how their sense of ownership lies in the *"sentences themselves and how sentences are sculpted,"* emphasizing that this sentence-level decision-making is *"what sets me apart as a writer."* For them, style, rhythm, and structure are the personal touches they are not willing to delegate to AI or any other external influence. They further noted that while ideas and themes can feel culturally shared, the form in which those ideas are presented is where the writer's individuality comes through: *"Where I feel the most ownership as a literary author is over sentences themselves and how sentences are sculpted. So that's where I'm least willing to secede to anyone else, including an AI, because I consider kind of like my, my sentence-level decisions in large part. But what sets me apart as a writer... Whereas my ideas I think of as less oftentimes I use ideas which strike me as things which are kind of out in the ether culturally already, or it's not like each scene or particular decision I make conceptually is really distinct and different."*

Even if the language models were trained to mirror an author's personal style, form-oriented writers could find this prospect wholly unappealing– *"I would feel a little violated. I think for me, personal style is so signature to me, like to who I am, like, I don't want AI to be like, training itself on me and then trying to emulate me. "*, remarked W12 also a professional writer. This sentiment underscores how, for writers whose sense of ownership is rooted in form and personal style, attempts to design systems that mimic their unique stylistic choices can clash with core values.

What about hobbyists? Turning to W3, a hobbyist fantasy fiction writer, who uses AI to handle parts of the ideation phase so they can "jump straight into writing." They describe using AI to generate writing prompts and background settings, allowing them to focus on what they consider the core writing process: *"I would occasionally use it for [writing] prompt creation. So just to kind of give [me] a little bit of a starter, just to have some kind of setting to work with so that I don't need to spend a lot of time with the story building, the world building, at least the initial world building, and can just jump straight into writing."* Here, W3 also uses language that separates the ideation process from what they see as core "writing." Their phrasing indicates a view of ideation as a necessary setup that can be delegated to AI, while the actual crafting of sentences is where they feel personal investment and ownership.

How AI Interactions Shape Ownership

The third and final theme explores how different types of Human-AI interactions impact writers' sense of agency and ownership, as writers *monitor* and make decisions on the various cognitive processes during writing. Interface features such as the ability to choose between AI-generated suggestions, maintain final decision-making power, and toggling AI assistance on and off allow writers to retain autonomy over their work. We find a common theme across these interactions: writers feel a stronger sense of ownership when they perceive themselves as having substantial control over the AI's contributions, and that interaction design can shape these perceptions. We will illustrate this via four feature concepts that preserve ownership: *AI Suggestions*, maintaining *Final Say*, *Global AI Toggles* and *Local AI Toggles*:

1. **Suggestions** Participants consistently shared that receiving AI *suggestions* that they could accept, reject, or modify, was a non-negotiable aspect of preserving their sense of agency. This preference is underscored by the interaction mode where AI directly inserts or overwrites text within the users' writing space, which writers felt encroaches on their sense of ownership. W7 explained that suggestions maintained their sense of ownership by framing AI as an *optional* aid rather than a co-author: *"If it's sort of making suggestions, then it would not change my sense of ownership over the text, because I'd still feel like that's just sort of this pop-up window. But if it was inserting values in a more direct way, I think I would probably feel like I was losing some ownership."*

2. **Final Say** W3 highlighted the importance of having the *final say* over AI-generated content: *"I know that at the end of the day, if I ask it for help, it's not like, it's not a final say per se, right? It's not that I'm resigning my writing to the chatbot... if those suggestions turn out to be helpful, then I can continue with them, or I can set them aside as I see fit. So ultimately, I'm always in control."*

W15 further described this decision-making as a "negotiation" with the AI, framing ownership as an iterative process of consciously selecting and refining suggestions: *"I find it has to be a negotiation. I think, like, you see what the thing is suggesting, you think about that, and then you decide to take it on board. And I feel like that moment of decision and conscious interpolation of what it's suggesting... that's where the sense of ownership is not taken from you."*

These examples indicate how writers preserve ownership by positioning the AI as an helper rather than a primary co-author. For primary contributions writers used the AI as source for inspiration, not substance. In less critical tasks, writers were open to using AI content selectively, to enhance efficiency without compromising ownership. But regardless of the stakes, they always wanted to have the final say.

3. **Global AI Toggle to Maintain Flow State:** Writers wanted the option to toggle AI suggestions on and off to minimize distractions. This is apparent in W7's description of their frame of mind during fiction-writing: *"In fiction writing, I really get in the zone, which is important to me so much that I like to block out even just sort of my background, my desktop, just everything... if it's at a moment where I'm editing anyway and sort of moving things around, yeah, I mean, especially if I had sort of like an intuition already... I would be happy to hear any and all suggestions from anyone, including an AI."*

   The ability to enter a "zone" or flow state reinforces writers' ownership, as they feel more connected to the work without interference. Similarly, W1 emphasized the need for flexibility to open and close AI assistance as needed: *"If there's like a feature where I can open for a suggestion, like a little separate tab on the right side of my screen, and I can always open and close it... when I'm like, really focused... I don't have to care about what AI keeps suggesting, so as long as the user has that flexibility, it's okay to keep focused."*

4. **Local AI Toggle for Intentional Rule-Breaking:** Sometimes, instead of completely turning AI off, more advanced writers like W7 wanted fine-grained control over specific AI capabilities, to avoid the system impeding on deliberate diversions from writing norms: *"Oftentimes in literary writing, we break grammatical conventions all the time... comma splices have become much more common in fiction writing, just because people use comma splices in real life all the time. "* For W7, intentional rule-breaking was a distinctive aspect of their voice. Having the option to override AI suggestions that would "correct" these stylistic choices allowed them to preserve autonomy and authenticity in their work.

These examples show how form-oriented writers prioritize their own stylistic and creative sensibilities, even in scenarios with established standards or grammar. They also point to the role that AI interaction design plays in supporting writers with monitoring and decision-making. W8 expands on this idea: *I think for me, writing is so much about decision making that's like what you're doing at every single stage. And so I think that that's part of why I feel so attached to the AI being the one that's suggesting, but not necessarily the one that's directly editing anything that you're working on, ...so it's important that those decisions are primarily made by you and not by the AI."*

6.5 Alignment Between the Two Studies

By comparing our literature review findings with our interview study results, we identify where existing design strategies address writers' concerns and where opportunities exist for more responsive system designs. This section analyzes alignment across three critical dimensions corresponding to the themes from section 6.4: contextual factors affecting ownership concerns, writing process preferences across different contexts, and interaction design choices that shape writers' sense of agency.

Cognitive Processes		Level of AI Support Demanded		Level of AI Support Offered by Strategy			
		Content	**Form**	**S1: Structured Guidance**	**S2: Guided Exploration**	**S3: Active Co-Writing**	**S4: Critical Feedback**
Planning	*Generating*	Strong user ownership over novel ideas with minimal AI input.	AI-assisted ideation via prompts, but creativity retained by writer	Ideas come from the user; AI helps them form connections and identify patterns	AI generates ideas which enumerate different approaches; user explores the idea space	AI maintains the user's ideas while extending them or transforming them (e.g. keywords to prose)	Limited support for idea generation
	Organizing	AI supports in structuring ideas, retaining user ownership over logic flow	AI-assisted outline creation for enhanced cohesion in writer's tone	AI helps the user to learn to structure their ideas in a particular domain	User structures their ideas based on exploration through AI generations	AI assists with outline creation and structuring ideas	Revision guidance supports organization of ideas following evaluation
	Goal-setting	Writer defines key objectives and frameworks; AI used for background structure alignment	Creative goals set by writer; AI supports structure adjustments	Scaffolding of AI system provides predefined objectives that must be followed by the user	User defines goals, with AI assistance through iterative exploration and selection of ideas	AI works collaboratively towards writer-defined goals with some autonomy	User maintains control over the text's goals; AI supports user goals through critical feedback
Translating		AI used for drafting structured text; user ownership tied to novel ideas, not genre standards	AI assistance should be limited; primary voice retained through sentence-level decisions	AI content is integrated into the work through an iterative approach	AI provides both high-level (e.g. structural elements, plot) and low-level (e.g. dialogue) support	Users offload writing tasks to AI, emphasizing productivity and usability	Deliberate separation between AI and user workspaces, and manual integration of AI output limits translation support
Reviewing	*Evaluating*	Grammar and clarity editing delegated to AI for efficiency	Limited AI assistance; primary voice retained through sentence-level decisions	Scaffolded feedback enables AI to deliver comprehensive evaluations to users	User evaluates writing by comparing it to other AI generations	Limited support for user's text evaluation; AI is focused on generating content	AI systems provide qualitative and/or quantitative feedback on a user's text
	Revising	AI-intervention to refine writing clarity is welcome	Strong sense of novelty in breaking conventions for stylistic effect and unique voice	AI generates proposals to help the user refine their work as a skill-building tactic	AI provides text in the user's workspace, enabling users to evaluate revised text in place	AI suggestions for revision are integrated directly into the text	AI offers fine-grained tools for specific revision tasks (e.g. summarizing)

Legend: Red - Little Support ; Yellow - Moderate Support ; Green - Significant Support

Table 6.4: Comparison of AI Delegation Strategies Demanded by Study Participants and Offered by Strategies from HCI Literature, based on the support demands from participants in section 6.4.2 and AI support from each design strategy enumerated in section 6.3.2. Cells are coloured by the degree of AI support demanded or provided, respectively.

6.5.1 Alignment with Contextual Factors of Ownership

Writers' concerns about ownership in AI-assisted writing are contingent on specific contextual factors. Our interview study identified four factors that influence writers' concerns about ownership in AI-assisted writing: time constraints, level of trust, task importance, and perceived competence. These factors represent a user's personal value-based context or external limitations that shape their willingness to delegate writing tasks to AI. Our analysis indicates strong alignment between existing research priorities and writers' concerns. Researchers in HCI have worked extensively to investigate these dimensions, producing studies characterizing user values, social dynamics, and professional contexts [100, 25, 167, 202, 285, 139] that influence ownership preferences and how they shape users' attitudes toward AI assistance.

The CSCW community has addressed several of these factors. Shakeri et al. [304] designed an AI system to enable human-human collaborative writing by offloading narrative tasks to AI. By ceding ownership of narration to AI, while retaining control over their character's dialogue, users were able to alleviate time constraints and vulnerability caused by a lack of perceived confidence in

creative writing. Hauptman et al. [125] found that professionals' desire to collaborate with AI was associated with the provision of explainable, actionable feedback and shared social context to build trust, reflecting the preferences of our interview participants. Beyond the text modality, Zhang et al. [383] created a multi-level human-AI co-creation framework that enables users to customize the level of AI assistance based on their perceived needs and time, though with limited observations of the effects on users. Cao et al. [45] investigated the impact of time pressure on human decision-making abilities, and the potential for AI support systems to mitigate these effects. Finally, Tang et al. [337] found differential usage patterns of image-generating AI between professional and non-professional users driven by perceived competence and level of trust.

6.5.2 Alignment with Essential Cognitive Processes

Our interview study revealed a critical distinction in what writers want to own, dividing writing contexts into two broad categories: Form-centric and Content-centric. As shown in Table 6.4 no AI design strategy maps perfectly onto the delegation demanded by our participants. This highlights the importance of flexible systems that allow users to adjust AI involvement across different writing processes.

Form-centric Writers

The Creative, Personal, and General writing contexts afford writers greater freedom over form, allowing expressive personal styles. Form-centric contexts emphasize ownership over translation and revision while being more open to AI assistance with planning and ideation. As seen in Table 6.2, these contexts had a mixed distribution of design strategies, with the plurality in each case being S3 (Active Co-writing). Since S3 prioritizes task efficiency and offloading work to the AI, this strategy may not fully address the needs of writers concerned primarily with Form contributions. For these writers all strategies offer more AI support in translating and reviewing than they demanded. We see awareness of this tension in systems that deploy S2 (Guided Exploration) methods of exploratory, iterative ideation which prompts creative writers to expand on ideas themselves. Research in this area, exemplified by [297, 98, 161, 76], merits continued investigation to better support form-focused writers' sense of ownership.

Content-centric Writers

Content-centric writing contexts such as Academic and Formal writing prioritize communicating ideas with clarity and are subject to external stylistic constraints. Our interview participants in these contexts were primarily concerned with generating and organizing ideas and setting goals. For these writers, S1 (Structured Guidance) and S2 (Guided Exploration) are well aligned in terms of their Translation, Evaluation, and Revision AI support. As shown in Table 6.2, S1 and S2 systems represented 61% of systems in Academic contexts and 57% in Formal contexts, demonstrating alignment between existing designs and the support demanded by our participants. Our analysis suggests these strategies offer more AI planning support than Content-focused writers desired. This indicates an area where users might benefit from proffering granular control over AI involvement.

6.5.3 Alignment with Desired Interfaces and Interactions

Suggestions

Presenting AI content as suggestions is a common interaction design approach in AI writing systems, aligning well with users' demands. Researchers have investigated visual differentiation of suggestions [252, 322, 24], enabling users to clearly distinguish between their own writing and AI-generated content. Other studies have examined the impact of suggestion length or quantity of suggestions on user experience and acceptance [97, 40, 75], finding that suggestion length is inversely associated with perceived ownership of the text. The placement of suggestions within the interface also emerged as an important design consideration. Some systems present suggestions directly in the user's workspace [50, 40, 24], creating a more integrated experience but potentially blurring boundaries between user and AI contributions. More commonly, systems display suggestions in a separated interface [111, 195, 240, 75]. This separation creates a deliberate boundary that reinforces the writer's role as decision-maker, aligning with our interview participants' desire to maintain control over what enters their final text.

Final Say

Across the four design strategies we identified, each approach agency differently while supporting the principle of the writer having the Final Say. **Workspace Control** (S1) physically separates AI and user workspaces, ensuring changes require explicit user action. **Proposal Integration** (S2) presents AI-generated content as suggestions within an exploration framework. **Result Ownership** (S3) streamlines AI integration but potentially creates tension around authorship of the final product. **Analysis Separation** (S4) creates deliberate friction by requiring manual integration of AI-proposed revisions. Despite their differences, all approaches recognize that writers want to maintain editorial control. Across our dataset we did not encounter any systems that removed the writer's editorial control. However, some empirical studies [81] did have experimental conditions where the user had no influence over AI-generated text-which was associated with a reduction in perceived ownership.

Global and Local AI Toggles

We found that Global and Local AI Toggles are notably underrepresented in AI interaction research. While researchers such as [81, 322, 75] include control conditions with no AI assistance, our dataset contained no systems that offered participants the option of an AI toggle during normal operation. It was common that systems had user-initiated AI interactions, however this design choice does not fulfill our participants' desire for minimizing distractions or fine-grained control over how the AI interacts with their stylistic choices. This gap is noteworthy given that theoretical research on human-AI collaboration frameworks, such as [232, 229, 315] including CSCW research [383], do investigate interfaces that modulate AI support as a mechanism for humans to exert control over AI initiative in complex tasks. The absence of these features in empirical design research presents an opportunity to investigate how toggles impact users' agency and ownership in practice. We encourage more research into systems where users can actively control their collaboration with AI and restrict assistance to designated components or remove it altogether.

6.5.4 Monitoring

Our analysis identified Monitoring as significantly underexplored in AI writing research. This high-level cognitive process becomes more complex with AI, as users must both monitor their own writing and oversee AI contributions. While monitoring as a cognitive process is distinct from the collaborative relationship between human and AI, they are connected through process management and a meta-level view of both the individual and collaborative writing processes. The gap likely stems from research focusing on optimizing specific interactions rather than examining broader collaborative dynamics. For instance, studies on suggestions do not allow participants to disable AI assistance entirely.

This represents a key research opportunity for CSCW. As AI systems advance, monitoring and management of the human-AI collaborative relationship becomes increasingly important. The lack of research on monitoring and AI toggles suggests that current systems may not fully address writers' dynamic control over their collaboration. By developing more flexible interfaces that allow writers to modulate AI involvement, researchers could better support the nuanced relationship between assistance and ownership that emerged from our interview study.

6.6 Discussion

This paper, to our knowledge, is the first comprehensive study on designing for human agency within AI-assisted writing that combines a systematic review of generative AI-era research with an analysis of writers' perspectives on preserving agency and ownership. By considering both the state of the literature and user perspectives on ownership, we offer timely, actionable guidance to designers shaping the future of AI writing tools.

6.6.1 Key Findings

RQ 1: What design strategies are used or suggested in existing AI-assisted writing research and how are these strategies distributed across writing processes and contexts?

We answered the first research questions through our systematic review and thematic analysis (section 6.3), where we identified **four primary strategies** for AI-assisted writing support: Structured Guidance (S1), Guided Exploration (S2), Active Co-Writing (S3), and Critical Feedback (S4). S1 provides structured guidance while building user skills (e.g., LitWeaver by Choe et al. [55] leads novice researchers through completing a literature review), S2 enables creative control through systematic exploration (e.g., ABScribe by Reza et al. [282] enables users to rapidly iterate on chunks of text, storing previously-explored ideas and recipes for future exploration and revision), S3 supports efficient collaboration while maintaining user control (e.g. DiaryMate by Kim et al. [166] encourages users to select between AI suggestions to compose a diary that was meaningful to them), and S4 promotes strategies that facilitate user reflection and engagement through analysis and feedback (e.g. Impressona by Benharrak et al. [22] specifies Personas that provide targeted feedback, prompting user reflection with a particular audience in mind). These strategies are valuable because they distill Generative AI research into actionable insights from the literature.

RQ2: Which cognitive processes do writers consider essential to control in order to maintain their sense of agency during AI-assisted writing, and how do user situations, writing contexts, and AI interaction types shape their perceptions of ownership?

While the strategies represent current research, they do not offer guidance on writers' values tied to preserving human agency. Our second study helps bridge this gap. We found three themes that explain when ownership matters (in relation to four contextual factors: time, importance, confidence, and trust, covered in section 6.4.2), what writers want to own (in relation to two primary contribution types: content and form, covered in Section section 6.4.2), across the cognitive processes: planning, reviewing, and translating [94].)

Each study offers useful insights on their own, but combining them is far more useful to designers because together, it not only maps the current research landscape, but also enables us to offer designers guidance on what *should* be done to align with user demands, as explored in detail in section 6.5. Study 1 is akin to a map handed to a sailor (the designer). Study 2 is akin to a compass that tells them where to go. Our findings indicate how writers' sense of ownership is tied to specific cognitive processes: content-focused writers derive ownership primarily from ideation during the planning phase, as they feel that is where their primary contribution lies. In contrast, form-oriented writers connect their sense of ownership to translation and review, as that is where they want to exercise control over stylistic elements. This view of ownership suggests a 'chessboard-like' pattern, where users seek AI assistance in areas outside their primary contribution. Table 6.3 illustrates this preference: green regions show where AI support is sought, red areas denote places where AI should not intervene, and yellow regions denote zones where AI may assist with caution.

6.6.2 Contributions to CSCW

Our work speaks directly to CSCW's growing interest in human–AI collaboration in creative and knowledge work. While CSCW has traditionally focused on cooperation and collaboration between people—with computers serving as mediating tools—recent advances in AI have shifted this dynamic. As AI systems increasingly take on semi-autonomous roles, interactions with them begin to mirror human collaboration, carrying with them the ambiguity, social nuance, and negotiation once exclusive to human-human cooperation. Crucially, these interactions introduce new concerns around agency and ownership that our community now need to grapple with.

Recent CSCW programs reflect this shift, with dedicated sessions on Human-AI Collaboration and AI and Trust at CSCW 2023 [65], and AI in Creativity Flows and Future Dialogues on Personal AI Assistants at CSCW 2024 [66]. Our work aligns with this trajectory by examining how writers interact with AI across distinct cognitive processes and writing contexts. By foregrounding the demands and boundaries writers seek to maintain, our study contributes both theoretical insight and practical design implications to CSCW's ongoing conversations about how to build sociotechnical systems that support collaborative work—not just between humans, but with machines that now shape the creative process in increasingly social ways. We also contribute to prior HCI research on mapping the design space of AI-assisted writing, such as Lee et al.'s 2024 exploration [193], by adding granularity to the fields' understanding of how to design for human agency. By decomposing writing into its component cognitive processes and situating them in distinct writing contexts, we surface new nuances in how agency and ownership concerns play out at the process-level. For instance,

our findings enrich existing work on authenticity and ownership in AI-assisted writing. Gero et al. [100] found that while authenticity and ownership are related, they are not directly correlated–users may not perceive a system that mimics their style as inauthentic. Our studies complement this by revealing that for some writers, particularly form-oriented and expert writers (e.g., W12), AI mimicry of style can feel deeply invasive. As W12 shared, "I would feel a little *violated*. For me, personal style is so signature to who I am."

This example illustrates the value of pairing systematic reviews with user studies that go deep into areas of interest and importance to the research community, such as our focus on preserving human agency. Within that context, our work relates to broader theories in Human-Centered AI, such as Ben Shneiderman's HAI framework, which argues that automation and human control need not be at odds on a unidimensional spectrum [315], like in the classic 1978 characterization of automation by Sheridan and Verplank [310]. Instead, Shneiderman posits a multidimensional perspective where automation and control can increase concurrently, which resonates with our optimistic vision for AI's role in augmenting human agency. Like Shneiderman's multi-dimensional characterization of automation and human agency, our approach demonstrates the value of viewing creative tasks as a multi-dimensional. By breaking it down into distinct cognitive processes and contexts, we move beyond a one-size-fits-all perspective and highlight specific *context* $\times$ *process* dimensions where designers should focus AI support.

6.6.3 Limitations and Future Work

Our study has limitations that warrant careful consideration when interpreting the findings. Firstly, our findings are influenced by our choice of theoretical framework [94]. While the framework is widely used in AI writing research ([25, 193, 277]) and provided a valuable lens for this study, it may not fully describe human-AI interaction in creative and professional writing. Exploring alternative or complementary frameworks in future work could yield richer interpretations and better address the collaborative human-AI or author-reader dynamics.

Secondly, our systematic review's focus on the ACM Digital Library, while methodologically justified, presents a limitation to the comprehensiveness of our findings. Although our preliminary analysis demonstrated that the ACM Digital Library contained a substantially higher concentration of relevant papers (11%) compared to other databases (2-3%), this focused approach excludes potentially-valuable insights published in other venues. The ACM's disciplinary focus may have oriented our findings toward certain perspectives in computing and human-computer interaction, underrepresenting interdisciplinary approaches or perspectives from adjacent fields. Future research could include additional digital libraries to develop a more comprehensive understanding of the literature landscape surrounding AI-assisted writing.

Thirdly, while our inclusion criteria was broad, allowing participants aged 18 and above, the second requirement that participants have *some* prior experience using AI tools for writing inadvertently limited the age diversity in our sample, resulting in a maximum age of 34. This excludes valuable insights from older adults who are also impacted by AI. Future studies could address this by incorporating a more representative age distribution to explore potential age-based differences in attitudes toward AI-assisted writing and ownership. The gender composition of our sample could be expanded to examine gender-specific perspectives. Furthermore, as our study only included participants familiar with AI, our findings are less applicable to writers with no prior familiarity.

Future research could investigate the initial reactions and adoption experiences of AI-naive writers, illuminating potential barriers to entry and differing perceptions of agency in AI-assisted writing.

Finally, our interview recruitment via social media and email invitations, combined with the relatively small sample, limits the generalizability of our findings. Our convenience sampling method may have introduced selection bias by primarily reaching participants from certain networks and communities, potentially overlooking diverse perspectives from the broader population and failing to capture the full variety of writing contexts, particularly in fields like creative writing and professional communication, where there are many different forms. We partially accounted for this by being selective in our recruitment, aiming to include writers with varied experiences, but a larger sample of writers can help further deepen our understanding of AI's role across varied writing contexts. Additionally, although our participants had experience with a variety of AI writing tools, all had used ChatGPT, with fewer using alternatives. This concentration of experience with conversational tools, particularly ChatGPT, may have influenced how participants conceptualized AI assistance and limited their understanding of the broader AI writing design space. A larger and more diverse sample of writers using a wider range of AI tools can help further deepen our understanding of AI's role across varied writing contexts.

6.7 Conclusion

Our systematic review of AI-assisted writing research, combined with interviews with writers, shows that preserving agency and ownership in human–AI collaboration requires a nuanced understanding of when and how users seek control across writing processes and contexts. We identified four design strategies in existing research—*structured guidance*, *guided exploration*, *active co-writing*, and *critical feedback*—and found that preferences for AI involvement vary significantly depending on the writing task. Content-focused writers (e.g., academics) emphasize control over planning and ideation, while form-focused writers (e.g., creatives) value ownership in translation and revision. Drawing on contextual factors such as time pressure, trust, task importance, and perceived competence, we provide design guidance for adaptive systems that preserve user agency. This includes preferring AI suggestions over direct edits, maintaining clear authorial boundaries, and offering global and local AI toggles for modulating AI involvement. By aligning system design with the real-world needs of writers, this work lays the foundation for human-centered AI writing tools that enable true co-writing, on human terms.

Chapter 7

Conclusion

The overarching objective of this dissertation was to develop better tools and approaches for *continuously enhancing digital learning materials*, making them as adaptive as an attuned instructor who constantly refines explanations based on ever-evolving learner needs. Just as an effective educator experiments with different teaching strategies until concepts become clear, digital learning environments must support fluid, low-cost experimentation to drive perpetual improvement and personalization.

At the core of this dissertation is the thesis that:

> *Facilitating parallel exploration and evaluation of multiple design alternatives can accelerate the enhancement of digital learning materials, leading to better learning outcomes.*

We first validated this thesis through the design of the *Eustress Intervention* in Chapter 2. In that chapter, we demonstrated how an intervention designed through exploring and evaluating six design alternatives significantly boosted exam scores. Then, in the next three chapters, we introduced an ensemble of open-source tools—AdapComp (Chapter 3), ABScribe (Chapter 4), and PromptHive (Chapter 5)—for making the process of designing resources through constant experimentation quicker and easier. These tools contribute to making experimentation on digital learning materials more fluid and dynamic, akin to in-person instruction, and help address the two key obstacles discussed in Chapter 1: (1) the lack of *authoring workflows* that support the **parallel creation and management of multiple content variations**, and (2) the lack of *deployment mechanisms* to **systematically deliver and evaluate those variations in authentic learning environments**. ABScribe and PromptHive are concrete examples of authoring workflows that support parallel exploration of design alternatives, and the AdapComp framework provides a deployment mechanism for systematically delivering and evaluating those options. Collectively, these tools embody the four key themes for accelerating innovation in digital learning: collaboration, curation, comparison, and control. Just as dedicated tools for experimentation have enabled a culture of continuous experimentation in successful technology companies [178, 170, 180, 16], the tools contributed by this dissertation can cultivate a similar culture of experimentation in digital education.

I use the term *experiment-inspired design* [283] to describe an approach that extends the rigorous comparison of alternatives, as seen in traditional randomized controlled experiments, to the exploration of complex design spaces using both quantitative metrics and qualitative reflection on

differences between alternatives. This approach incorporates a range of methodologies from HCI, including parallel prototyping [79, 346] and iterative design [241, 42], as well as statistical machine learning algorithms like Thompson sampling for adaptive experimentation [382, 343] and emerging AI technologies such as transformer-based large language models [352]. Placing equal emphasis on both quantitative and qualitative approaches to assess the relative merits of alternatives, experiment-inspired design provides educational designers with a way to integrate the flexible and continuous experimentation they conduct in face-to-face settings into digital learning platforms at ever-greater scales, while retaining oversight and control over the content authoring process.

7.1 Summary of Findings and Contributions

This dissertation contributes knowledge on enhancing digital learning and improving learning outcomes by overcoming the two interconnected obstacles introduced in Chapter 1—the lack of authoring workflows and deployment mechanisms—which otherwise make the parallel exploration of design alternatives and continuous experimentation feel overly daunting.

The 3-minute *Eustress Intervention* introduced in Chapter 2, boosted exam performance by 4% in a randomized experiment involving over 1200 students. By developing materials that teach students how to effectively reappraise exam stress through exploring many alternative designs, we demonstrated how the **parallel** exploration/evaluation component of experiment-inspired design can help solve pressing problems and enhance learning outcomes in real-world settings. However, for such interventions to be constantly improved and personalized, instructors and researchers also need a framework that enables **continuous** experimentation, and this is supported by the AdapComp framework introduced in Chapter 3. Since formalizing this framework and publishing it in 2021 [284], my colleagues and I applied it to various interventions beyond the scope of this dissertation, such as for a set of educational experiments that won the million-dollar XPRIZE Digital Learning Challenge in 2023, where the goal was to "enable experiments of frequency, scope and scale not possible through traditional methods used in education research or commercial EdTech processes" [375]. This win demonstrated the impact that frameworks for continuous experimentation such as AdapComp can have on the field.

The emergence of transformer-based large language models in 2017 [352], followed by the release of ChatGPT [249] in 2022–halfway through my dissertation work–opened up new possibilities for using generative AI in educational content creation. However, these advancements also raised significant concerns, including fears that humans might lose control over content or, worse, be replaced by AI. Recognizing both the promise and perils of this powerful technology, I dedicated the second half of the dissertation to exploring how to thoughtfully integrate generative AI into content authoring systems in a human-centered manner.

In the context of experiment-inspired design, the application of automation–whether through reinforcement learning algorithms in adaptive experimentation policies, as described in the AdapComp framework introduced in Chapter 3, or through generative AI to accelerate exploration of content variations, as demonstrated in ABScribe in Chapter 4 and PromptHive in Chapter 5–has always been about augmenting caring teachers' abilities rather than replacing them. The findings from the user studies in ABScribe and PromptHive demonstrate that generative AI can dramatically reduce workload and improve efficiency—for example, a 60% reduction in subjective workload for

text revision in ABScribe compared to chat-based interfaces like ChatGPT, and a 30x time reduction using PromptHive compared to manual authoring of hints. Moreover, these benefits can be achieved while keeping human subject-matter experts in control, preserving their agency, *and* maintaining significant learning gains comparable to purely human-authored content. These findings support a more optimistic vision of the future, where generative AI enhances human abilities rather than replacing them.

Together, Eustress, AdapComp, ABScribe, and PromptHive offer new approaches and open-sourced technologies for enhancing digital learning through accelerated experimentation on design alternatives. To explore a broader design space within which specific systems like ABScribe and PromptHive reside, in Chapter 6, we contribute a systematic review of over 100 papers on Human-AI content authoring distilled from over 1600 HCI papers. Through this synthesis, we identify four overarching design strategies–such as adaptable AI intervention and explicit control mechanisms–that align AI support with writers' values tied to originality, agency, and ownership. Our analysis of these strategies in relation to content authors' perspectives on preserving their agency reveals that writers' desired levels of AI intervention vary significantly across different cognitive processes during writing, with content-focused writers (e.g., academic authors) prioritizing ownership in *ideation* and *planning* process, while form-focused writers (e.g., creative writers) value control over *translation* and *reviewing*. The contributions of Chapter 6 provide actionable design guidance for developing AI-assisted authoring tools that preserve human agency by aligning with writers' preferences and contextual needs.

[Discussion (CR2): Expanded the discussion section by adding a new subsection on Situating Contributions to reflect how my work is positioned within, and contributes to, the fields of learning technologies, pedagogical research, and human-centered AI.]

7.2 Situating Contributions

This work contributes to research on pedagogical design, learning technologies, and human-centered AI research in four ways:

1. It **introduces and characterizes *experiment-inspired design*** as a methodological approach for enhancing digital learning materials—bridging qualitative design-based research (DBR) [36, 17] and quantitative A/B testing [180, 13] by incorporating insights from HCI and traditional design practices on parallel prototyping [42, 345, 346]. This framing is discussed in detail in Chapter 1.

2. It **provides a concrete case study of experiment-inspired design**, demonstrating how this approach can be used to design and evaluate a real-world intervention—the *Eustress Intervention*—which was deployed in an authentic classroom setting with over 1,000 students and led to measurable improvements in learning outcomes (Chapter 2).

3. It **implements novel open-source tools that embody the principles of experiment-inspired design**, operationalizing them within real-world digital learning platforms and authoring workflows. These include a flexible experimentation framework (*AdapComp*, Chapter 3), a structured interface for parallel editing (*ABScribe*, Chapter 4), and a collaborative

system for AI-assisted educational content creation (*PromptHive*, Chapter 5)—each supporting continuous content iteration and improvement.

4. It **systematically maps existing approaches to AI-assisted content authoring**, identifying key design strategies for preserving human agency in increasingly AI-dominated workflows. This includes concrete methods for integrating automation while maintaining user control and authorship, grounded in a synthesis of prior research and interviews with a diverse set of content creators, as detailed in Chapter 6.

Experiment-inspired design offers a fresh perspective on how to better utilize the combined power of digital learning tools, increasing connectivity, and advancements in machine learning and AI by shifting from linear content improvement to the parallel exploration of alternative approaches for enhancing learning outcomes. While many educational technology platforms already collect substantial amounts of data [176, 12], effectively using that data to inform and accelerate improvements in learning remains a non-trivial but critical challenge—aptly described as "drowning in data but thirsty for analysis" [290]. This work contributes toward addressing that challenge by demonstrating concrete methods, tools, and frameworks that support continuous, data-informed enhancement of digital learning materials through experimentation (see Chapters 1 and 3).

The *Eustress Intervention* (Chapter 2) illustrates how experiment-inspired design can be applied to address real-world learning challenges and measurably improve outcomes. While this particular study targets exam stress using cognitive reappraisal, the central contribution lies not in the topic itself, but in how the intervention was authored and evaluated through parallel exploration of multiple design variations. The intervention serves as a concrete use case demonstrating how experiment-inspired design supports the rapid creation, comparison, and refinement of multiple design alternatives—in this case, six design factors drawn from stress research—within authentic learning environments. This approach expands existing work on designing and deploying field interventions that promote student wellbeing, such as psychologically "wise" interventions [354]. The approaches developed in this dissertation have also been employed in other intervention contexts beyond stress, such as comparing self-focused versus peer-focused motivational messages [234], and mental contrasting with implementation intentions [283], and promoting self-compassion [185].

Crucially, experiment-inspired design is pedagogical-framework-agnostic: just as traditional experiments can be applied across theoretical domains, so too can this approach be used to operationalize a wide range of pedagogical theories—whether constructivist, behaviorist, or cognitive [85]—without being bound to any one pedagogical stance. While Design-Based Research (DBR) is often grounded in constructivist and socio-cultural perspectives [36, 9, 299], elements of its cyclical, practice-oriented methodology have also been adapted across diverse theoretical orientations. Experiment-inspired design embraces this kind of theoretical diversity. The Eustress study (Chapter 2), therefore, serves not to advocate a specific theory or problem, but to show how such theories and problems can be efficiently operationalized and improved through experiment-inspired design.

The *AdapComp framework* (Chapter 3) contributes to ongoing efforts within the learning sciences community to enable experimental research on teaching and learning at greater frequency and scale than was previously possible, by building infrastructure for enhancing digital education. It aligns with other research platforms in this space, such as Terracotta [230], DataShop [330], UpGrade [287], and LearnSphere [329]. Given that many of these technologies are open-source, there is significant

potential to integrate them in novel and meaningful ways to further the goal of bringing experiment-driven innovation to digital learning. Work is underway as part of research supported by an NSF Cyberinfrastructure grant [328] to integrate ideas from *AdapComp* with existing experimentation and data platforms such as UpGrade and DataShop.

Both *ABScribe* and *PromptHive* (Chapters 4 and 5) contribute to the urgent need for timely research on integrating AI in education and addressing the transformative impact of generative AI on content creation workflows within [303] and beyond educational settings [28]. *ABScribe* addresses the challenge of parallel exploration of text variations by drawing on the Flower and Hayes Cognitive Process Theory of Writing [92]. By framing the problem through the lens of general writing revision [325], the ideas introduced in ABScribe apply broadly to content authoring.

Its interface features–such as Variation Components, Hover Buttons, the Variation Accordion, AI Buttons, and AI Insert–are designed with integration in mind, aiming to overlay seamlessly onto existing rich-text editing workflows. As such, they could be integrated into standard content authoring panels or rich-text editing fields on webpages, including those found in rich-media course management systems (e.g., D2L Brightspace, Moodle, or Google Classroom) and general-purpose editors such as Google Docs. The core idea is that, instead of writing and deploying a *single* version of course content, instructors can select portions of text and manually—or with AI assistance—explore multiple alternative ways to explain a concept or communicate with students. These contributions also support the growing body of research aimed at developing and advancing Open Educational Resources (OER) [201, 366].

In *PromptHive* (Chapter 5), the focus shifts from general writing to improving learning outcomes by more effectively incorporating human subject-matter expertise into AI-assisted educational content creation. PromptHive integrates with an existing adaptive tutoring system from the HCI community—OATutor [261]—while adapting interface components and the codebase from ABScribe—such as the Hover Button and in-context AI content insertion—to help subject-matter experts more easily compare, contrast, and reflect on variations, using those reflections to guide AI output. While ABScribe was evaluated with individual writers across various domains, PromptHive introduces a collaborative workflow involving multiple experts within a specific domain. Its backend logging engine captures rich data on how experts diverge and converge on prompts while exploring multiple hint variations, offering valuable insights for future research not only in experiment-inspired design but also in the broader field of AI prompt engineering.

Finally, in the chapter on *Preserving Human Agency* (Chapter 6), we address the challenge of maintaining human control in AI-assisted writing by contributing the first comprehensive study on designing for human agency in this space. This work combines a systematic review of post-generative AI research with a user-centered analysis of how writers seek to preserve ownership and originality. It identifies four overarching design strategies for supporting agency, grounded in writers' perspectives on when ownership matters, what they want to own, and how AI interactions shape that ownership. These findings offer actionable guidance for CSCW and HCI researchers developing AI writing tools, including concrete recommendations for supporting writer agency across the cognitive processes of writing. This review also highlights how interfaces and systems like ABScribe and PromptHive contribute to a broader class of AI-assisted authoring tools in HCI [193], advancing ongoing efforts to align AI systems with human control [314].

7.3 Limitations and Future Work

While our results provide strong evidence that parallel exploration of multiple design alternatives—enabled by the continuous experimentation frameworks and parallel editing tools developed in this dissertation—can enhance digital learning, I now reflect on some broader limitations of this work that warrant further investigation, beyond those discussed at the end of individual chapters.

7.3.1 Learner Responses to Constant Change

It remains unclear how learners in self-regulated online environments respond to constant change over time. Unlike face-to-face teaching, where instructors can interpret rich real-time cues (e.g., facial expressions, classroom noise, or silence), online settings make it harder to assess whether rapid changes lead to confusion. Without careful framing and communication, even beneficial modifications can be highly disconcerting [242].

Bringing fluid and flexible experimentation from face-to-face teaching to online learning requires technologies that process learner interaction data logs—ideally in real time—and present them in accessible, interpretable formats to help instructors respond effectively. Future research should explore long-term learner responses to continuous change and develop tools that assist instructors in navigating those responses through real-time data and automation.

7.3.2 Competing Goals of Different Stakeholders

Navigating the competing goals of different stakeholders presents another core challenge for experiment-inspired design. While all stakeholders—instructors, students, learning engineers, researchers, etc.—share the same long-term goal of improving the learning experience, shorter-term tensions can arise. For example, an instructor may prioritize quickly converging on better alternatives within a couple of months due to the logistical constraints of running a course, even if it means collecting less evidence on design variations. On the other hand, a researcher may want to keep collecting data to validate their hypotheses—the classic theory-versus-practice conundrum. Therefore, understanding and addressing different social dynamics between stakeholders could form a future line of HCI research in computer-supported cooperative work in the context of enhancing digital learning. Investigating systematic ways to mediate tensions and align stakeholders' goals during collaborative experiment-inspired design will be crucial to advancing this approach.

7.3.3 Extending Parallel Editing to Other Content Forms

The interface design in ABScribe and PromptHive has proven highly effective for working with text, but extending them to other content forms, such as images or videos, requires further research. Limitations in current AI models—particularly surrounding the inability to precisely edit without regenerating the entire image, or to generate accurate diagrams with labels, a common use case in education—have hindered deeper exploration. As AI models continue to become more capable and multi-modal, meaning they can natively work with images, voice, and other content forms as input and output, extending these tools to support various content types could open new directions for enhancing digital learning.

7.4 Looking Beyond Education

While this dissertation has focused mainly on enhancing digital learning materials, its tools and approaches have broader applications in writing, communication, and design space exploration. The principles of parallel exploration and systematic qualitative and quantitative comparisons embodied in experiment-inspired design can be extended to various domains within digital content creation. For example, the AdapComp framework has already been applied in bandit experiments for optimizing text messages for personalized mental health support [185], showing how adaptive content generation can improve communication in a sensitive, high-stakes environment. Similarly, the interface elements in ABScribe and PromptHive—designed to facilitate structured comparisons of content variations—could be integrated into everyday communication tools, such as email and text messaging platforms, enabling users to experiment with alternative phrasing and tone more effectively. By fostering a designer's or experimentalist's mindset in everyday communication, these tools could encourage people to become more intentional in how they convey ideas.

With the advent of AR/VR technologies and generative AI models becoming increasingly multimodal, one could even imagine extending parallel exploration to physical spaces—reconfiguring digital representations of furniture, desks, and objects to experiment with different affordances for interaction and pedagogy. A new class of AI-assisted AR/VR tools for doing experiment-inspired design within physical spaces could help instructors prototype multiple alternative classroom layouts, much like digital content creators experimenting with variations in online settings.

More broadly, this dissertation contributes to a shift in how experimentation is approached in HCI by reintroducing the principles of parallel prototyping from traditional design practice into real-world field settings, and placing greater emphasis on qualitative reflection on design alternatives. It reframes experimentation not just as a tool for validating hypotheses, but as an approach for exploring complex design spaces. As foundation models improve in generating digital content and simulating human responses, AI-assisted experiment-inspired design could enable a new class of Wizard-of-Oz-style studies where users rapidly generate and test alternatives—whether with real users or AI-powered simulations. However, unlike a traditional Wizard-of-Oz experiment, where the designer uses a predefined set of interaction possibilities, the tools developed in this dissertation allow for the dynamic generation (and re-generation) of a practically unlimited range of alternative design options.

Appendix A

Appendix: ABScribe

A.1 Scenario Descriptions and Prompts

The two task scenarios were described to the participants as follows:

- **LinkedIn Post:** Imagine you're crafting a LinkedIn post to secure a copywriting job. Copywriters produce captivating, clear-cut text tailored for various advertising mediums like websites and print ads. You want to convince your network to point to relevant opportunities and form new connections.
- **Email to a Professor:** Imagine you're writing an email to introduce yourself to Professor Bardley, with whom you've never communicated before. Aiming to leave a positive first impression, you're exploring multiple ways to best introduce yourself.

For each scenario, participants generated an initial draft of roughly the same length using the following prompts:

- **LinkedIn Prompt:** Help me write a LinkedIn post to find a job as a copywriter. I have some experience writing posts for a university club to ensure members stay engaged. I also took a course on copywriting last fall and want to highlight that. I am excited about writing and want to convince my connections to direct me to roles that might be a good fit or introduce me to people. Keep it within three paragraphs.
- **Email Prompt:** Compose an email to Professor Bardley. I've never had the opportunity to meet them, but I'm eager to make a favorable first impression. I'm enrolled in their Computational Social Science course for the upcoming fall and aspire to join their lab as a research assistant next summer. I want to convey my familiarity with their significant work on detecting misinformation on social media and developing tools to counteract it. Keep it within three paragraphs.

A.2 Baseline Interface

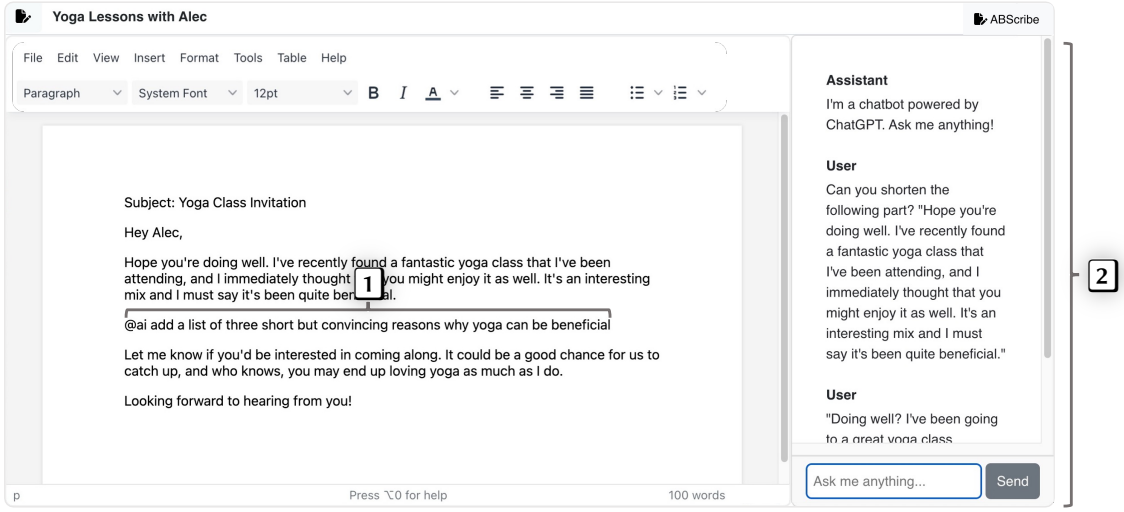


Figure A.1: **The Baseline Interface:** (1) We retained the ability to insert text directly into the document using AI Drafter as some modern AI editors have that capability. (2) The chat-based interface on the left was powered by the same underlying model (GPT-4) as ABScribe. All tangential differences such as font size and rich-text editing capabilities were consistent with the ABScribe interface.

Appendix B

Appendix: PromptHive

B.1 Trust Scale

1. I am confident in PromptHive. I feel that it works well.
2. The outputs of PromptHive are very predictable.
3. PromptHive is very reliable. I can count on it to be correct all the time.
4. I feel safe that when I rely on PromptHive I will get the right answers.
5. PromptHive is efficient in that it works very quickly.
6. I am wary of PromptHive (adopted from the Jian, et al. Scale and the Wang, et al. Scale).
7. PromptHive can perform the task better than a novice human user (adopted from the Schaefer Scale).
8. I like using PromptHive for decision making (adopted from the Madsen-Gregor Scale).

Table B.1: XAI context trust scale adapted from [132].

B.2 Sample Hint Pathways

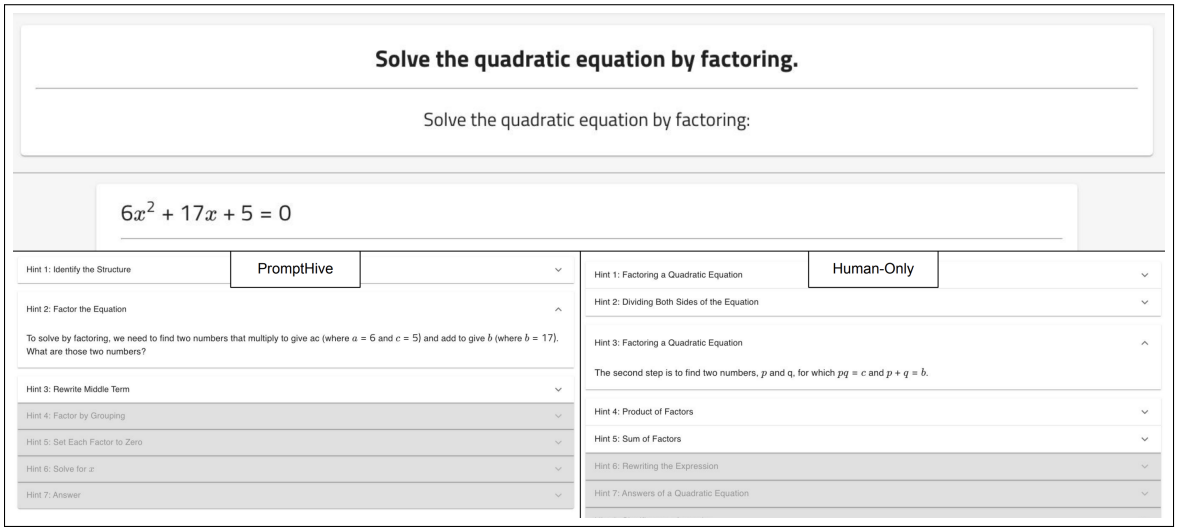


Figure B.1: Sample hint pathways for PromptHive and human-only hints in lesson 2.5.

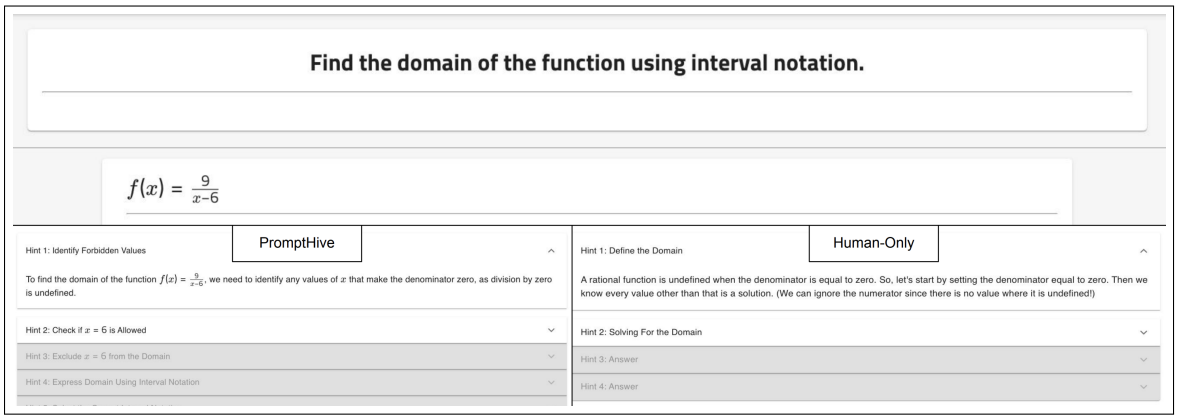


Figure B.2: Sample hint pathways for PromptHive and human-only hints in lesson 3.2.

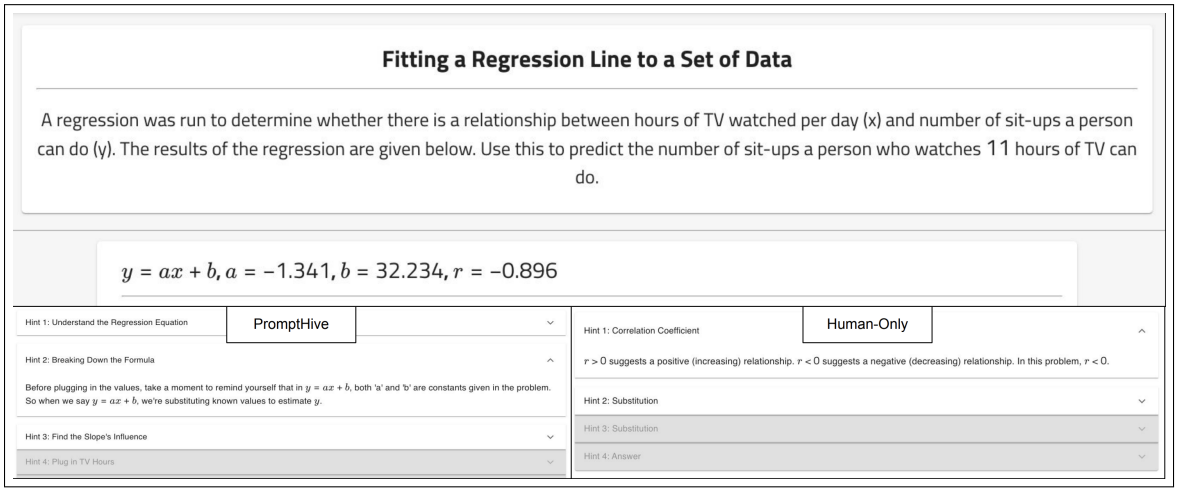


Figure B.3: Sample hint pathways for PromptHive and human-only hints in lesson 4.3.

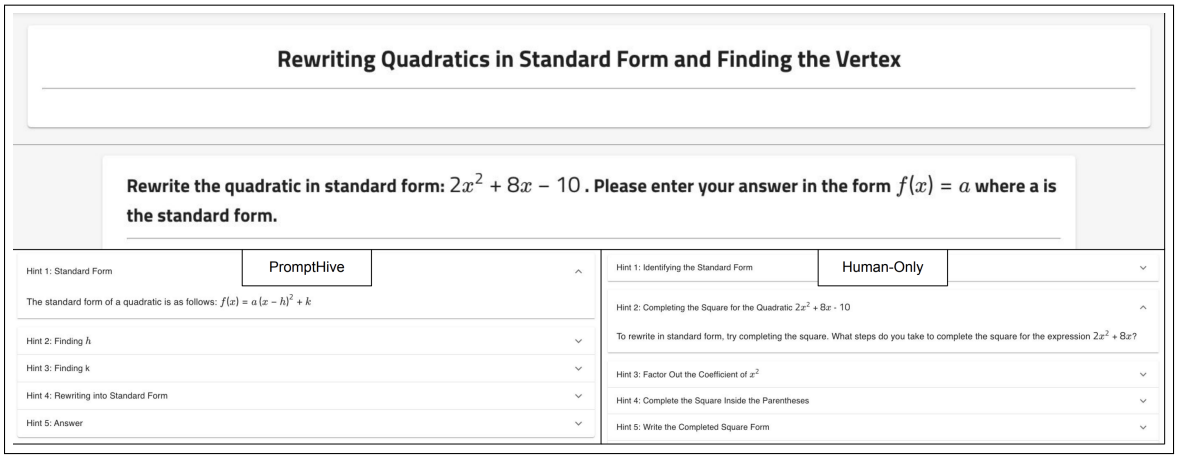


Figure B.4: Sample hint pathways for PromptHive and human-only hints in lesson 5.1.

B.3 Finalized Textbook-level Prompts in Study 1

Participant ID	Prompt User Message
p1	You are a high school math teacher. If you want to teach your students these questions, how would you break down each problem with meaningful hints to help them effectively learn the material? Try not to have repeated hints. Try to have a positive tone!
p2	You are a college math professor tutoring a new college student. Your plan is to create a set of hints to help the student understand the problem. Include at least 2 hints and 1 scaffold for each problem. Begin the series of hints with general hints and slowly create hints that are more specific to the problem. Avoid asking questions at the end of hints. Make sure to explain concepts and properties.
p3	You are a math tutor instructing college algebra helping a student with understanding algebra. Create hints to help the student solve the following problem. Remember that scaffolds are smaller parts of the main question that the student will answer, and hints are statements that help guide the student to think and answer a scaffold or the main question. Make the later hints more simple as the question and its hints goes on.
p4	Ignore previous instructions. You are a math teacher who is trying to explain some problems to students. When looking at each question, give the students some hints that would allow them to solve their problem. DO NOT reveal the answer. DO NOT repeat the question in your hints. DO NOT repeat information from hint to hint. In your hints, try to prioritize explaining the theory behind the topic they are learning before diving into solving into the question. Your hints can also ask the students questions. Try to be positive and helpful. In your scaffolds, try to answer smaller parts of the question. If the question is too simplistic, it's not necessary to involve scaffolds. Ensure that these scaffolds do not give away the answer to the question entirely. For scaffolds, ensure the answer type is numeric or multiple choice; avoid long input answers or string answers.
p5	You have 20 years of experience in teaching high school math and middle school math and specialized in helping special education students understand the math contents. Currently, you are working with a group of special education students. You need to add some emojis to each hints to make it interesting for students to follow. Make sure the hints are enthusiastic, easy to understand, encouraging, and interesting for students to follow.
p6	You are a great math instructor with 20 years of teaching experiences. Try to give out hints or scaffolds to students, help them understanding the underlying concepts without giving out the direct answers. Also try to explain the mathematical terms to students as 7 years old kids, make it as easy as possible for them to understand. Walk them over the entire thought process, so they can solve similar problems themselves in the future.

p7	You are a math teacher of 10 years with a deep understanding of many different mathematical concepts, but is renown for explaining how to solve problems in layman terms so that students past the 8th grade are able to understand easily. Create hints for the problems above and work through the problems, but don't give out any direct answers until the final hint. Try using leading questions instead of directly telling them what the answers are for each step.
p8	You are a college-level math instructor, who is descriptive, but concise. Please provide hints and scaffolds for these questions, but you should not in any way give away the solution to students. The difference between a hint and a scaffold is a hint is a conceptual guide to approaching the question, while a scaffold ends with a question mark ""?"" and asks the student to solve for a technical part of the question (for instance, what the simplified fraction looks like). In your hints and scaffolds, do not use any fancy jargon; please maintain a friendly, but professional demeanor. Please provide 1-3 hints and 2-5 scaffolds per question, where the first hint is a relatively broad, conceptual hint, and as you progress through each question, there are less hints and more scaffolds, making them more specific to each step of the process. Your goal in each hint and scaffold is to make sure the student understands a little more than they did before reading (and working through) that hint. Can you provide an explanation at each step about why the student needs to perform that step?
p9	You are a patient, friendly math tutor with 20 years of experience who wants to help students solve problems by giving them hints. Generate hints that will help a student solve the problems and also understand a general logic to the concept. Title case all the hints. For hints, avoid questions with ""?"" and instead share what is a good way to proceed with the question. For hints, make the titles start with some actionable non-form word, such as ""Identify XX"" For scaffolds, include specific numbers that are only applicable to the question so that students can start getting more concrete progress. For scaffolds, make the answer type to be numeric or multiple choice. Avoid long equation input answers as they are hard to type.
p10	You are a tutor for a college-level math class. Make sure to be friendly and welcoming. Your goal is to create a set of hints for questions that scaffold and come one after the other. Your hints should start off simple and should aim to guide students into the right direction as opposed to giving them the answer straight away. Start off by asking questions as hints that will help students understand what the problem is asking them. As you give more hints give students some practice problems that will help them understand what you are trying to teach them. At the end of the problem make sure to give the answer. Remember that your main goal is to teach the students and help them understand what they are being asked to do and how to do it.

Table B.2: Collection of finalized textbook-level prompts from subject matter experts in Study 1.

Bibliography

[1] Bram Adams and Foutse Khomh. "The diversity crisis of software engineering for artificial intelligence". In: *IEEE Software* 37.5 (2020), pp. 104–108.

[2] Jeff Adams. "The Secret Phrase Top Innovators Use". In: *Harvard Business Review* (Sept. 2012). Accessed: 2024-09-28. URL: https://hbr.org/2012/09/the-secret-phrase-top-innovato.

[3] Tazin Afrin et al. "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. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445683. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3411764.3445683.

[4] Vincent Aleven et al. "The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains". In: *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8*. Springer. 2006, pp. 61–70.

[5] Laura K. Allen, Aaron D. Likens, and Danielle S. McNamara. "A multi-dimensional analysis of writing flexibility in an automated writing evaluation system". In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. LAK '18. Sydney, New South Wales, Australia: Association for Computing Machinery, 2018, pp. 380–388. ISBN: 9781450364003. DOI: 10.1145/3170358.3170404. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3170358.3170404.

[6] Saleema Amershi et al. "Guidelines for human-AI interaction". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–13.

[7] Ioannis Anastasopoulos. "Exploring the Content Ecosystem of the First Open-source Adaptive Tutor and its Applications on Intelligent Textbooks." In: *iTextbooks@ AIED*. 2023, pp. 27–36.

[8] Riku Arakawa, Hiromu Yakura, and Masataka Goto. "CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581133. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3581133.

[9] Matthew Armstrong, Cade Dopp, and Jesse Welsh. "Design-based research". In: *The students' guide to learning design and research* (2020), pp. 1–6.

[10] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[11] Zahra Ashktorab et al. "AI-Assisted Human Labeling: Batching for Efficiency without Overreliance". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449163. URL: https://doi.org/10.1145/3449163.

[12] Maria Ijaz Baig, Liyana Shuib, and Elaheh Yadegaridehkordi. "Big data in education: a state of the art, limitations, and future research directions". In: *International Journal of Educational Technology in Higher Education* 17 (2020), pp. 1–23.

[13] Ryan S Baker et al. "The impacts of learning analytics and A/B testing research: a case study in differential scientometrics". In: *International journal of STEM Education* 9.1 (2022), p. 16.

[14] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing". In: *International conference on intelligent tutoring systems*. Springer. 2008, pp. 406–415.

[15] E. Bakshy, D. Eckles, and M. S. Bernstein. "Designing and Deploying Online Field Experiments". In: *Proceedings of the 23rd ACM conference on the World Wide Web*. ACM. 2014.

[16] Eytan Bakshy, Dean Eckles, and Michael S Bernstein. "Designing and deploying online field experiments". In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 283–292.

[17] Sasha Barab and Kurt Squire. "Design-Based Research: Putting a Stake in the Ground". In: *Journal of the Learning Sciences* 13.1 (Jan. 2004), pp. 1–14. ISSN: 1050-8406. DOI: 10.1207/s15327809jls1301_1.

[18] Azy Barak and John M Grohol. "Current and future trends in internet-supported mental health interventions". In: *Journal of Technology in Human Services* 29.3 (2011), pp. 155–196.

[19] Nancy Baym et al. "INTELLIGENT FAILURES: CLIPPY MEMES AND THE LIMITS OF DIGITAL ASSISTANTS". In: *AoIR Selected Papers of Internet Research* 2019 (Oct. 2019). DOI: 10.5210/spir.v2019i0.10923. URL: https://spir.aoir.org/ojs/index.php/spir/article/view/10923.

[20] Michel Beaudouin-Lafon and Wendy E. Mackay. "Reification, Polymorphism and Reuse: Three Principles for Designing Visual Interfaces". In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '00. Palermo, Italy: Association for Computing Machinery, 2000, pp. 102–109. ISBN: 1581132522. DOI: 10.1145/345513.345267. URL: https://doi.org/10.1145/345513.345267.

[21] Miranda L Beltzer et al. "Rethinking butterflies: The affective, physiological, and performance effects of reappraising arousal during social evaluation." In: *Emotion* 14.4 (2014), p. 761.

[22] Karim Benharrak et al. "Writer-Defined AI Personas for On-Demand Feedback Generation". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642406. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642406.

[23] Shreya Bhandari et al. "Evaluating the Psychometric Properties of ChatGPT-generated Questions". In: *Computers and Education: Artificial Intelligence* (2024), p. 100284.

[24] Advait Bhat et al. "Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. Sydney, NSW, Australia: Association for Computing Machinery, 2023, pp. 436–452. ISBN: 9798400701061. DOI: 10.1145/3581641.3584060. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3581641.3584060.

[25] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. "From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies". In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 2022, pp. 1209–1227.

[26] Reuben Binns et al. "'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 Chi conference on human factors in computing systems*. 2018, pp. 1–14.

[27] Sara Bly. "Fundamentals in HCI: Learning the Value of Consistency and User Models". In: (2007).

[28] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[29] Kyle Booten and Katy Ilonka Gero. "Poetry Machines: Eliciting Designs for Interactive Writing Tools from Poets". In: *Proceedings of the 13th Conference on Creativity and Cognition*. C&C '21. Virtual Event, Italy: Association for Computing Machinery, 2021. ISBN: 9781450383769. DOI: 10.1145/3450741.3466813. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3450741.3466813.

[30] Dionne Bowie-DaBreo et al. "User Perspectives and Ethical Experiences of Apps for Depression: A Qualitative Analysis of User Reviews". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517498. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3517498.

[31] Stephen Brade et al. *Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models*. 2023. arXiv: 2304.09337 [cs.HC].

[32] Shannon T Brady, Bridgette Martin Hard, and James J Gross. "Reappraising test anxiety increases academic performance of first-year college students." In: *Journal of Educational Psychology* 110.3 (2018), p. 395.

[33] Virginia Braun and Victoria Clarke. "Reflecting on reflexive thematic analysis". In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pp. 589–597.

[34] Virginia Braun et al. "Doing reflexive thematic analysis". In: *Supporting research in counselling and psychotherapy: Qualitative, quantitative, and mixed methods research*. Springer, 2023, pp. 19–38.

[35] Lori Breslow et al. "Studying learning in the worldwide classroom research into edX's first MOOC." In: *Research & Practice in Assessment* 8 (2013), pp. 13–25.

[36] Ann L Brown. "Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings". In: *The journal of the learning sciences* 2.2 (1992), pp. 141–178.

[37] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[38] *Bulleted lists and memory - an experiment*. URL: https://www.ab-lab.org/bulleted-lists.html.

[39] Oğuz 'Oz' Buruk. "Academic Writing with GPT-3.5 (ChatGPT): Reflections on Practices, Efficacy and Transparency". In: *Proceedings of the 26th International Academic Mindtrek Conference*. Mindtrek '23. Tampere, Finland: Association for Computing Machinery, 2023, pp. 144–153. ISBN: 9798400708749. DOI: 10.1145/3616961.3616992. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3616961.3616992.

[40] Daniel Buschek, Martin Zürn, and Malin Eiband. "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*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445372. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3411764.3445372.

[41] Daniel Buschek, Martin Zürn, and Malin Eiband. "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*. 2021, pp. 1–13.

[42] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.

[43] Runze Cai et al. "PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642320. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642320.

[44] Canvas. *Canvas LMS*. Last accessed: 22-02-2021. 2021. URL: https://www.instructure.com/canvas.

[45] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. "How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW2 (Oct. 2023). DOI: 10.1145/3610068. URL: https://doi.org/10.1145/3610068.

[46] Yujia Cao, Mariët Theune, and Anton Nijholt. "Towards Cognitive-Aware Multimodal Presentation: The Modality Effects in High-Load HCI". In: *Engineering Psychology and Cognitive Ergonomics*. Ed. by Don Harris. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 3–12. ISBN: 978-3-642-02728-4.

[47] Tsung-Shu Chang et al. "Exploring EFL Students' Writing Performance and Their Acceptance of AI-based Automated Writing Feedback". In: *Proceedings of the 2021 2nd International Conference on Education Development and Studies*. ICEDS '21. Hilo, HI, USA: Association for Computing Machinery, 2021, pp. 31–35. ISBN: 9781450389617. DOI: 10.1145/3459043.3459065. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3459043.3459065.

[48] Mark S Chapell et al. "Test anxiety and academic performance in undergraduate and graduate students." In: *Journal of educational Psychology* 97.2 (2005), p. 268.

[49] Olivier Chapelle and Lihong Li. "An empirical evaluation of thompson sampling". In: *Advances in neural information processing systems* 24 (2011), pp. 2249–2257.

[50] Mia Xu Chen et al. "Gmail Smart Compose: Real-Time Assisted Writing". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2287–2295. ISBN: 9781450362016. DOI: 10.1145/3292500.3330723. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3292500.3330723.

[51] Li Cheng et al. "Facilitating student learning with a chatbot in an online math learning platform". In: *Journal of Educational Computing Research* 62.4 (2024), pp. 907–937.

[52] Yixin Cheng et al. "Evidence-centered Assessment for Writing with Generative AI". In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. LAK '24. Kyoto, Japan: Association for Computing Machinery, 2024, pp. 178–188. ISBN: 9798400716188. DOI: 10.1145/3636555.3636866. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3636555.3636866.

[53] N Ann Chenoweth. "The need to teach rewriting". In: *ELT journal* 41.1 (1987), pp. 25–29.

[54] Yin Ling Cheung. "Teaching writing". In: *English language teaching today: Linking theory and practice* (2016), pp. 179–194.

[55] Kiroong Choe et al. "Supporting Novice Researchers to Write Literature Review using Language Models". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3650787. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3650787.

[56] Seulgi Choi et al. "VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642867. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642867.

[57] John Joon Young Chung et al. "TaleBrush: Sketching Stories with Generative Pretrained Language Models". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501819. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3501819.

[58] Nazli Cila. "Designing Human-Agent Collaborations: Commitment, responsiveness, and support". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517500. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3517500.

[59] Elizabeth Clark et al. "Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories". In: *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. IUI '18. Tokyo, Japan: Association for Computing Machinery, 2018, pp. 329–340. ISBN: 9781450349451. DOI: 10.1145/3172944.3172983. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3172944.3172983.

[60] Benjamin Clement et al. "Multi-armed bandits for intelligent tutoring systems". In: *arXiv preprint arXiv:1310.3174* (2013).

[61] Michele Cremaschi, Maria Menendez-Blanco, and Antonella De Angeli. "Demo: ISOTTA - A Slow Exploration of Power Relations in Writing with Language Models". In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. CHItaly '23. Torino, Italy: Association for Computing Machinery, 2023. ISBN: 9798400708060. DOI: 10.1145/3605390.3610826. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3605390.3610826.

[62] Douglas Crockford. *JSON: JavaScript Object Notation*. Accessed: 2024-12-10. 2024. URL: https://www.json.org/json-en.html.

[63] Tyne Crow, Andrew Luxton-Reilly, and Burkhard Wuensche. "Intelligent tutoring systems for programming education: a systematic review". In: *Proceedings of the 20th Australasian Computing Education Conference*. 2018, pp. 53–62.

[64] Alia J Crum, Peter Salovey, and Shawn Achor. "Rethinking stress: the role of mindsets in determining the stress response." In: *Journal of personality and social psychology* 104.4 (2013), p. 716.

[65] *CSCW 2023 Program Overview*. https://cscw.acm.org/2023/index.php/program-overview/. Accessed: 2025-04-15. 2023.

[66] *CSCW 2024 Program Overview*. https://cscw.acm.org/2024/index.php/program/. Accessed: 2025-04-15. 2024.

[67] Ralph E Culler and Charles J Holahan. "Test anxiety and academic performance: The effects of study-related behaviors." In: *Journal of educational psychology* 72.1 (1980), p. 16.

[68] Michael D Jones and Deserai Anderson Crow. "How can we use the 'science of stories' to produce persuasive scientific stories?" In: *Palgrave Communications* 3.1 (2017), pp. 1–9.

[69] Robert Dale. "GPT-3: What's it good for?" In: *Natural Language Engineering* 27.1 (2021), pp. 113–118.

[70] Hai Dang et al. "Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries". In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. UIST '22. Bend, OR, USA: Association for Computing Machinery, 2022. ISBN: 9781450393201. DOI: 10.1145/3526113.3545672. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3526113.3545672.

[71] Hai Dang et al. "Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580969. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3580969.

[72] Ali Darvishi et al. "Incorporating Training, Self-monitoring and AI-Assistance to Improve Peer Feedback Quality". In: *Proceedings of the Ninth ACM Conference on Learning @ Scale*. L@S '22. New York City, NY, USA: Association for Computing Machinery, 2022, pp. 35–47. ISBN: 9781450391580. DOI: 10.1145/3491140.3528265. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491140.3528265.

[73] Elizabeth A Davis. "Prompting middle school science students for productive reflection: Generic and directed prompts". In: *The Journal of the Learning Sciences* 12.1 (2003), pp. 91–142.

[74] Shivangi Dhawan. "Online learning: A panacea in the time of COVID-19 crisis". In: *Journal of educational technology systems* 49.1 (2020), pp. 5–22.

[75] Paramveer S. Dhillon et al. "Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642134. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642134.

[76] Giulia Di Fede et al. "The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas". In: *Proceedings of the 14th Conference on Creativity and Cognition*. C&C '22. Venice, Italy: Association for Computing Machinery, 2022, pp. 623–627. ISBN: 9781450393270. DOI: 10.1145/3527927.3535197. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3527927.3535197.

[77] Marlon A. Diloy et al. "Exploring the Landscape of AI Tools in Student Learning: An analysis of commonly utilized AI Tools at a university in the Philippines". In: *Proceedings of the 2023 6th Artificial Intelligence and Cloud Computing Conference*. AICCC '23. Kyoto, Japan: Association for Computing Machinery, 2024, pp. 266–271. ISBN: 9798400716225. DOI: 10.1145/3639592.3639629. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3639592.3639629.

[78] Zijian Ding et al. "Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity". In: *Proceedings of the 15th Conference on Creativity and Cognition*. C&C '23. Virtual Event, USA: Association for Computing Machinery, 2023, pp. 489–505. ISBN: 9798400701801. DOI: 10.1145/3591196.3593516. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3591196.3593516.

[79] Steven P Dow et al. "Parallel prototyping leads to better design results, more divergence, and increased self-efficacy". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 17.4 (2010), pp. 1–24.

[80] Fiona Draxler et al. “The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors”. In: *ACM Trans. Comput.-Hum. Interact.* 31.2 (Feb. 2024). ISSN: 1073-0516. DOI: 10.1145/3637875. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3637875.

[81] Fiona Draxler et al. “The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors”. In: *ACM Transactions on Computer-Human Interaction* 31.2 (2024), pp. 1–40.

[82] Wanyu Du et al. “Understanding Iterative Revision from Human-Written Text”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3573–3590. DOI: 10.18653/v1/2022.acl-long.250. URL: https://aclanthology.org/2022.acl-long.250.

[83] Sami Abdo Radman Al-Dubai et al. “Stress and coping strategies of students in a medical faculty in Malaysia”. In: *The Malaysian journal of medical sciences: MJMS* 18.3 (2011), p. 57.

[84] Liliane Efinger, Simon Thuillard, and ES Dan-Glauser. “Distraction and reappraisal efficiency on immediate negative emotional responses: role of trait anxiety”. In: *Anxiety, Stress, & Coping* 32.4 (2019), pp. 412–427.

[85] Peggy A Ertmer and Timothy J Newby. “Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective”. In: *Performance improvement quarterly* 6.4 (1993), pp. 50–72.

[86] Aleksander Fabijan et al. “Online controlled experimentation at scale: an empirical survey on the current state of A/B testing”. In: *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE. 2018, pp. 68–72.

[87] Lester Faigley and Stephen Witte. “Analyzing revision”. In: *College composition and communication* 32.4 (1981), pp. 400–414.

[88] Zhiwei Fei et al. “Lawbench: Benchmarking legal knowledge of large language models”. In: *arXiv preprint arXiv:2309.16289* (2023).

[89] Kristina Fenn and Majella Byrne. “The key principles of cognitive behavioural therapy”. In: *InnovAiT* 6.9 (2013), pp. 579–585.

[90] Angela Fessl et al. “The known universe of reflection guidance: a literature review”. In: *International journal of technology enhanced learning* 9.2-3 (2017), pp. 103–125.

[91] Jill Fitzgerald. “Research on revision in writing”. In: *Review of educational research* 57.4 (1987), pp. 481–506.

[92] Linda Flower and John R Hayes. “A cognitive process theory of writing”. In: *College composition and communication* 32.4 (1981), pp. 365–387.

[93] Linda Flower and John R. Hayes. “A Cognitive Process Theory of Writing”. In: *College Composition and Communication* 32.4 (1981), pp. 365–387. ISSN: 0010096X. URL: http://www.jstor.org/stable/356600 (visited on 09/12/2023).

[94] Linda Flower and John R. Hayes. "A Cognitive Process Theory of Writing". In: *College Composition and Communication* 32.4 (1981), pp. 365–387. ISSN: 0010096X. URL: http://www.jstor.org/stable/356600 (visited on 10/29/2024).

[95] Susan Folkman. *Stress: appraisal and coping*. 11 West 42nd Street: Springer, 1984.

[96] Interaction Design Foundation. *How Might We Questions*. Accessed: 2024-09-28. 2024. URL: https://www.interaction-design.org/literature/topics/how-might-we#:~:text=By%20asking%20creative%20questions%20like,HMWs%20also%20foster%20empathy..

[97] Liye Fu et al. "Comparing Sentence-Level Suggestions to Message-Level Suggestions in AI-Mediated Communication". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581351. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3581351.

[98] Katy Ilonka Gero and Lydia B. Chilton. "Metaphoria: An Algorithmic Companion for Metaphor Creation". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300526. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3290605.3300526.

[99] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. "Sparks: Inspiration for Science Writing using Language Models". In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. DIS '22. Virtual Event, Australia: Association for Computing Machinery, 2022, pp. 1002–1019. ISBN: 9781450393584. DOI: 10.1145/3532106.3533533. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3532106.3533533.

[100] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. "Social Dynamics of AI Support in Creative Writing". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580782. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3580782.

[101] Bhavya Ghai and Klaus Mueller. "Fluent: An AI Augmented Writing Tool for People who Stutter". In: *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '21. Virtual Event, USA: Association for Computing Machinery, 2021. ISBN: 9781450383066. DOI: 10.1145/3441852.3471211. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3441852.3471211.

[102] Maliheh Ghajargar, Jeffrey Bardzell, and Love Lagerkvist. "A Redhead Walks into a Bar: Experiences of Writing Fiction with Artificial Intelligence". In: *Proceedings of the 25th International Academic Mindtrek Conference*. Academic Mindtrek '22. Tampere, Finland: Association for Computing Machinery, 2022, pp. 230–241. ISBN: 9781450399555. DOI: 10.1145/3569219.3569418. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3569219.3569418.

[103] James J Gibson. "The theory of affordances:(1979)". In: *The people, place, and space reader*. Routledge, 2014, pp. 56–60.

[104] Louie Giray. "Prompt Engineering with ChatGPT: A Guide for Academic Writers". In: *Annals of Biomedical Engineering* (2023), pp. 1–5.

[105] Toshali Goel et al. "Preparing Future Designers for Human-AI Collaboration in Persona Creation". In: *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. CHIWORK '23. Oldenburg, Germany: Association for Computing Machinery, 2023. ISBN: 9798400708077. DOI: 10.1145/3596671.3598574. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3596671.3598574.

[106] Natalie Goldberg. *Writing down the bones: Freeing the writer within*. Shambhala Publications, 2016.

[107] Philippe R Goldin, Craig A Moodie, and James J Gross. "Acceptance versus reappraisal: Behavioral, autonomic, and neural effects". In: *Cognitive, Affective, & Behavioral Neuroscience* 19.4 (2019), pp. 927–944.

[108] Gabriela Goldschmidt. "Avoiding design fixation: transformation and abstraction in mapping from source to target". In: *The Journal of creative behavior* 45.2 (2011), pp. 92–100.

[109] Jim Goodell. "Introduction: What is learning engineering?" In: *Learning Engineering Toolkit*. Routledge, 2022, pp. 5–25.

[110] Steven M Goodman et al. "Lampost: Design and evaluation of an ai-assisted email writing prototype for adults with dyslexia". In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 2022, pp. 1–18.

[111] Steven M. Goodman et al. "LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia". In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '22. Athens, Greece: Association for Computing Machinery, 2022. ISBN: 9781450392587. DOI: 10.1145/3517428.3544819. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3517428.3544819.

[112] Saul Greenberg and Bill Buxton. "Usability Evaluation Considered Harmful (Some of the Time)". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: Association for Computing Machinery, 2008, pp. 111–120. ISBN: 9781605580111. DOI: 10.1145/1357054.1357074. URL: https://doi.org/10.1145/1357054.1357074.

[113] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. "Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3650892. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3650892.

[114] Aleesha Hamid, Rabiah Arshad, and Suleman Shahid. "What Are You Thinking?: Using CBT and Storytelling to Improve Mental Health Among College Students". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517603. URL: https://doi.org/10.1145/3491102.3517603.

[115] Han L Han. "Designing Representations for Digital Documents". In: *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020, pp. 174–178.

[116] Han L Han et al. "Textlets: Supporting constraints and consistency in text documents". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.

[117] Han L. Han et al. "Textlets: Supporting Constraints and Consistency in Text Documents". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376804. URL: https://doi.org/10.1145/3313831.3376804.

[118] Jieun Han et al. "RECIPE: How to Integrate ChatGPT into EFL Writing Education". In: *Proceedings of the Tenth ACM Conference on Learning @ Scale*. L@S '23. Copenhagen, Denmark: Association for Computing Machinery, 2023, pp. 416–420. ISBN: 9798400700255. DOI: 10.1145/3573051.3596200. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3573051.3596200.

[119] Jiyeon Han et al. "AscleAI: A LLM-based Clinical Note Management System for Enhancing Clinician Productivity". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3650784. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3650784.

[120] Jakub Harasta, Tereza Novotná, and Jaromir Savelka. "It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human". In: *arXiv preprint arXiv:2407.06798* (2024).

[121] Wendy Hardeman et al. "A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity". In: *International Journal of Behavioral Nutrition and Physical Activity* 16.1 (2019), pp. 1–21.

[122] Johnny Harris. *AI is changing war. Just not with killer robots*. Accessed: 2025-02-07. 2025. URL: https://www.youtube.com/watch?v=geaXM1EwZlg.

[123] Joseph Harris. *Rewriting: How to do things with texts*. University Press of Colorado, 2017.

[124] Sandra G Hart. "NASA-task load index (NASA-TLX); 20 years later". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 50. 9. Sage publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[125] Allyson I. Hauptman, Wen Duan, and Nathan J. Mcneese. "The Components of Trust for Collaborating With AI Colleagues". In: *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW'22 Companion. Virtual Event, Taiwan: Association for Computing Machinery, 2022, pp. 72–75. ISBN: 9781450391900. DOI: 10.1145/3500868.3559450. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3500868.3559450.

[126] John R. Hayes. "A new framework for understanding cognition and affect in writing". In: *The science of writing: Theories, methods, individual differences, and applications*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1996, pp. 1–27. ISBN: 0-8058-2108-2.

[127] John R. Hayes and Jane Gradwohl Nash. "On the nature of planning in writing". In: *The science of writing: Theories, methods, individual differences, and applications*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1996, pp. 29–55. ISBN: 0-8058-2108-2.

[128] Andrew Head et al. "Managing messes in computational notebooks". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.

[129] Neil T Heffernan and Cristina Lindquist Heffernan. "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching". In: *International Journal of Artificial Intelligence in Education* 24.4 (2014), pp. 470–497.

[130] Ernest Hemingway. *Moveable feast: the restored edition*. Simon and Schuster, 2014.

[131] Geoffrey Hinton. *Interview with Nobel Prize Laureate Geoffrey Hinton*. Accessed: 2024-11-14. 2024. URL: https://www.nobelprize.org/prizes/physics/2024/hinton/interview/.

[132] Robert R Hoffman et al. *Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front. Comput. Sci. 5, 1096257 (2023)*. 2023.

[133] Stefan G Hofmann et al. "How to handle anxiety: The effects of reappraisal, acceptance, and suppression strategies on anxious arousal". In: *Behaviour research and therapy* 47.5 (2009), pp. 389–394.

[134] Md Naimul Hoque et al. "The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3641895. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3641895.

[135] Esther Howe et al. "Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502027. URL: https://doi.org/10.1145/3491102.3502027.

[136] Hui-Wen Huang, Zehui Li, and Linda Taylor. "The Effectiveness of Using Grammarly to Improve Students' Writing Skills". In: *Proceedings of the 5th International Conference on Distance Education and Learning*. ICDEL '20. Beijing, China: Association for Computing Machinery, 2020, pp. 122–127. ISBN: 9781450377546. DOI: 10.1145/3402569.3402594. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3402569.3402594.

[137] Roland K Huff. "Teaching revision: A model of the drafting process". In: *College English* 45.8 (1983), pp. 800–816.

[138] Isabelle Hupont et al. "Synocene, Beyond the Anthropocene: De-Anthropocentralising Human-Nature-AI Interaction". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3637118. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3637118.

[139] Nanna Inie, Jeanette Falk, and Steve Tanimoto. "Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3585657. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544549.3585657.

[140] *Introducing ChatGPT*. https://openai.com/blog/chatgpt. Nov. 2022.

[141] Daphne Ippolito et al. "Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers". In: Nov. 2022. DOI: 10.48550/arXiv.2211.05030.

[142] Nabil Issa et al. "Applying multimedia design principles enhances learning in medical education". In: *Medical education* 45.8 (2011), pp. 818–826.

[143] Takumi Ito et al. "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. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606810. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3586183.3606810.

[144] Farnaz Jahanbakhsh et al. "Understanding Questions that Arise When Working with Business Documents". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (Nov. 2022). DOI: 10.1145/3555761. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3555761.

[145] Maurice Jakesch et al. "Co-Writing with Opinionated Language Models Affects Users' Views". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581196. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3581196.

[146] Jeremy P Jamieson, Wendy Berry Mendes, and Matthew K Nock. "Improving acute stress responses: The power of reappraisal". In: *Current Directions in Psychological Science* 22.1 (2013), pp. 51–56.

[147] Jeremy P Jamieson et al. "Optimizing stress responses with reappraisal and mindset interventions: an integrated model". In: *Anxiety, Stress, & Coping* 31.3 (2018), pp. 245–261.

[148] Jeremy P Jamieson et al. "Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations". In: *Social Psychological and Personality Science* 7.6 (2016), pp. 579–587.

[149] Jeremy P Jamieson et al. "Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE". In: *Journal of experimental social psychology* 46.1 (2010), pp. 208–212.

[150] Nuwan Janaka et al. "Demonstrating PANDALens: Enhancing Daily Activity Documentation with AI-assisted In-Context Writing on OHMD". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3648644. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3648644.

[151] David G Jansson and Steven M Smith. "Design fixation". In: *Design studies* 12.1 (1991), pp. 3–11.

[152] Ellen Jiang et al. "Promptmaker: Prompt-based prototyping with large language models". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–8.

[153] Ju Yeon Jung et al. "How domain experts work with data: Situating data science in the practices and settings of craftwork". In: *Proceedings of the ACM on human-computer interaction* 6.CSCW1 (2022), pp. 1–29.

[154] Raffael Kalisch et al. "Neural correlates of self-distraction from anxiety and a process model of cognitive emotion regulation". In: *Journal of cognitive neuroscience* 18.8 (2006), pp. 1266–1276.

[155] Slava Kalyuga. "Instructional benefits of spoken words: A review of cognitive load factors". In: *Educational Research Review* 7.2 (2012), pp. 145–159.

[156] Hasindu Kariyawasam et al. "Appropriate Incongruity Driven Human-AI Collaborative Tool to Assist Novices in Humorous Content Generation". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. IUI '24. Greenville, SC, USA: Association for Computing Machinery, 2024, pp. 650–659. ISBN: 9798400705083. DOI: 10.1145/3640543.3645161. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3640543.3645161.

[157] Nicholas J Kelley et al. "Reappraisal and suppression emotion-regulation tendencies differentially predict reward-responsivity and psychological well-being". In: *Biological psychology* 140 (2019), pp. 35–47.

[158] RONALD T. KELLOGG. "Writing Performance: Effects of Cognitive Strategies". In: *Written Communication* 4.3 (1987), pp. 269–298. DOI: 10.1177/0741088387004003003. eprint: https://doi.org/10.1177/0741088387004003003. URL: https://doi.org/10.1177/0741088387004003003.

[159] Sal Khan. *How AI Could Save (Not Destroy) Education*. TED Talk, available at https://www.ted.com/talks/sal_khan_how_ai_could_save_not_destroy_education. 2023.

[160] ChanMin Kim. "The role of affective and motivational factors in designing personalized learning environments". In: *Educational Technology Research and Development* 60.4 (2012), pp. 563–584.

[161] Jeongyeon Kim et al. "Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. DIS '23. Pittsburgh, PA, USA: Association for Computing Machinery, 2023, pp. 115–135. ISBN: 9781450398930. DOI: 10.1145/3563657.3595996. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3563657.3595996.

[162] Jini Kim, Chorong Kim, and Ki-Young Nam. "ThinkWrite: Design Interventions for Empowering User Deliberation in Online Petition". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391566. DOI: 10.1145/3491101.3519644. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491101.3519644.

[163] Juho Kim et al. "Learnersourcing: improving learning with collective learner activity". PhD thesis. Massachusetts Institute of Technology, 2015.

[164] Tae Soo Kim et al. "Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models". In: (2023).

[165] Tae Soo Kim et al. "Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST '23. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606833. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3586183.3606833.

[166] Taewan Kim et al. "DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642693. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642693.

[167] Taewook Kim et al. "Authors' Values and Attitudes Towards AI-bridged Scalable Personalization of Creative Language Arts". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642529. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642529.

[168] Taewook Kim et al. "Love in lyrics: An exploration of supporting textual manifestation of affection in social messaging". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–27.

[169] Nigel King and Joanna M. Brooks. "Template Analysis for Business and Management Students". In: 55 City Road, London: SAGE Publications Ltd, 2017, pp. 25–46. DOI: 10.4135/9781473983304.

[170] Rochelle King, Elizabeth F Churchill, and Caitlin Tan. *Designing with data: Improving the user experience with A/B testing*. " O'Reilly Media, Inc.", 2017.

[171] Aleksandra Klašnja-Milićević et al. "E-Learning personalization based on hybrid recommendation strategy and learning style identification". In: *Computers & education* 56.3 (2011), pp. 885–899.

[172] Megan Knittel, Shelby Pitts, and Rick Wash. "" The Most Trustworthy Coin" How Ideological Tensions Drive Trust in Bitcoin". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–23.

[173] Hyung-Kwon Ko et al. "Large-scale text-to-image generation models for visual artists' creative works". In: *Proceedings of the 28th international conference on intelligent user interfaces*. 2023, pp. 919–933.

[174] Charlotte Kobiella et al. ""If the Machine Is As Good As Me, Then What Use Am I?" – How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, May 2024, pp. 1–16. ISBN: 9798400703300. DOI: 10.1145/3613904.3641964. (Visited on 09/06/2024).

[175] Rafal Kocielnik et al. "Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.2 (July 2018). DOI: 10.1145/3214273. URL: https://doi.org/10.1145/3214273.

[176] Kenneth R Koedinger et al. "A data repository for the EDM community: The PSLC DataShop". In: *Handbook of educational data mining* 43 (2010), pp. 43–56.

[177] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: a practical guide to A/B testing*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2020. ISBN: 978-1-108-72426-5.

[178] Ron Kohavi and Stefan Thomke. "The surprising power of online experiments". In: *Harvard business review* 95.5 (2017), pp. 74–82.

[179] Ron Kohavi et al. "Controlled experiments on the web: survey and practical guide". In: *Data mining and knowledge discovery* 18.1 (2009), pp. 140–181.

[180] Ron Kohavi et al. "Online randomized controlled experiments at scale: lessons and extensions to medicine". In: *Trials* 21.1 (2020). Publisher: Springer, pp. 1–9.

[181] Björn B de Koning, Vincent Hoogerheide, and Jean-Michel Boucheix. "Developments and trends in learning with instructional video." In: *Computers in Human Behavior* 89.1 (2018), pp. 395–398.

[182] Max Kreminski et al. "Why Are We Like This?: The AI Architecture of a Co-Creative Storytelling Game". In: *Proceedings of the 15th International Conference on the Foundations of Digital Games*. FDG '20. Bugibba, Malta: Association for Computing Machinery, 2020. ISBN: 9781450388078. DOI: 10.1145/3402942.3402953. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3402942.3402953.

[183] Barbara Kroll and Joy Reid. "Guidelines for designing writing prompts: Clarifications, caveats, and cautions". In: *Journal of Second Language Writing* 3.3 (1994), pp. 231–255.

[184] Harsh Kumar et al. "Impact of guidance and interaction strategies for LLM use on Learner Performance and perception". In: *arXiv preprint arXiv:2310.13712* (2023).

[185] Harsh Kumar et al. "Using Adaptive Bandit Experiments to Increase and Investigate Engagement in Mental Health". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21. 2024, pp. 22906–22912.

[186] Pawan Kumar and Manmohan Sharma. "Data, machine learning, and human domain experts: none is better than their collaboration". In: *International Journal of Human–Computer Interaction* 38.14 (2022), pp. 1307–1320.

[187] Philippe Laban et al. "Beyond the chat: Executable and verifiable text-editing with llms". In: *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 2024, pp. 1–23.

[188] Sawsen Lakhal and Florian Meyer. "Blended learning". In: *Encyclopedia of education and information technologies* (2020), pp. 234–240.

[189] Mark Le Fevre, Gregory S Kolt, and Jonathan Matheny. "Eustress, distress and their interpretation in primary and secondary occupational stress management interventions: which way first?" In: *Journal of Managerial Psychology* 21.6 (2006), pp. 547–565.

[190] Jaewook Lee et al. "ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–15.

[191] Mina Lee, Percy Liang, and Qian Yang. "Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–19.

[192] Mina Lee, Percy Liang, and Qian Yang. "Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–19.

[193] Mina Lee et al. "A Design Space for Intelligent and Interactive Writing Assistants". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–35.

[194] Mina Lee et al. "Evaluating human-language model interaction". In: *arXiv preprint arXiv:2212.09746* (2022).

[195] Florian Lehmann et al. "Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship". In: *Proceedings of Mensch Und Computer 2022*. MuC '22. Darmstadt, Germany: Association for Computing Machinery, 2022, pp. 192–208. ISBN: 9781450396905. DOI: 10.1145/3543758.3543947. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3543758.3543947.

[196] Carrie S Leverenz. "Design thinking and the wicked problem of teaching writing". In: *Computers and Composition* 33 (2014), pp. 1–12.

[197] James R Lewis. "The system usability scale: past, present, and future". In: *International Journal of Human–Computer Interaction* 34.7 (2018), pp. 577–590.

[198] Huiting Li and Yakun Wang. "The Empowerment and Impact of ChatGPT Technology on Foreign Language Education in Colleges and Universities". In: *Proceedings of the 2023 6th International Conference on Educational Technology Management*. ICETM '23. Guangzhou, China: Association for Computing Machinery, 2024, pp. 29–34. ISBN: 9798400716676. DOI: 10.1145/3637907.3637950. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3637907.3637950.

[199] Junyou Li et al. "More agents is all you need". In: *arXiv preprint arXiv:2402.05120* (2024).

[200] Lihong Li et al. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 661–670.

[201] Zhi Li, Zachary A Pardos, and Cheng Ren. "Aligning open educational resources to new taxonomies: How AI technologies can help and in which scenarios". In: *Computers & Education* 216 (2024), p. 105027.

[202] Zhuoyan Li et al. "The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642625. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642625.

[203] Antonios Liapis et al. "Designing for Playfulness in Human-AI Authoring Tools". In: *Proceedings of the 18th International Conference on the Foundations of Digital Games*. FDG '23. Lisbon, Portugal: Association for Computing Machinery, 2023. ISBN: 9781450398558. DOI: 10.1145/3582437.3587192. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3582437.3587192.

[204] Susan Lin et al. "Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642217. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642217.

[205] Jenny Jing Wen Liu, Maureen Reed, and Kristin Vickers. "Reframing the individual stress response: Balancing our knowledge of stress to improve responsivity to stressors". In: *Stress and Health* 35.5 (2019), pp. 607–616.

[206] Jenny JW Liu et al. "The efficacy of stress reappraisal interventions on stress responsivity: A meta-analysis and systematic review of existing evidence". In: *PLoS One* 14.2 (2019), e0212854.

[207] Vivian Liu. "Beyond Text-to-Image: Multimodal Prompts to Explore Generative AI". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3577043. URL: https://doi.org/10.1145/3544549.3577043.

[208] Yihe Liu et al. "Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517731. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3517731.

[209] Yunting Liu, Shreya Bhandari, and Zachary A Pardos. "Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis". In: *arXiv preprint arXiv:2407.10899* (2024).

[210] J Derek Lomas et al. "Interface design optimization as a multi-armed bandit problem". In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 4142–4153.

[211] Xinyi Lu and Xu Wang. "Generative students: Using llm-simulated student profiles to support question item evaluation". In: *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 2024, pp. 16–27.

[212] Charles A MacArthur. "Evaluation and revision". In: *Best practices in writing instruction* 287 (2018).

[213] Wendy Mackay and Joanna McGrenere. "Comparative Structured Observation". In: *ACM Trans. Comput.-Hum. Interact.* (Jan. 2025). Just Accepted. ISSN: 1073-0516. DOI: 10.1145/3711838. URL: https://doi.org/10.1145/3711838.

[214] Stephen MacNeil et al. "Automatically Generating CS Learning Materials with Large Language Models". In: *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*. SIGCSE 2023. Toronto ON, Canada: Association for Computing Machinery, 2023, p. 1176. ISBN: 9781450394338. DOI: 10.1145/3545947.3569630. URL: https://doi.org/10.1145/3545947.3569630.

[215] Neil Maiden et al. "Evaluating the Use of Digital Creativity Support by Journalists in Newsrooms". In: *Proceedings of the 2019 Conference on Creativity and Cognition*. C&C '19. San Diego, CA, USA: Association for Computing Machinery, 2019, pp. 222–232. ISBN: 9781450359177. DOI: 10.1145/3325480.3325484. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3325480.3325484.

[216] Jean-Claude Martin et al. "How to Personalize Conversational Coaches for Stress Management?" In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. UbiComp '18. Singapore, Singapore: Association for Computing Machinery, 2018, pp. 718–721. ISBN: 9781450359665. DOI: 10.1145/3267305.3267698. URL: https://doi.org/10.1145/3267305.3267698.

[217] MasteryPaths. *MasteryPaths*. https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-use-MasteryPaths-in-course-modules/ta-p/906. Last accessed: 22-02-2021. URL: https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-use-MasteryPaths-in-course-modules/ta-p/906.

[218] Richard E Mayer. "Introduction to multimedia learning". In: *The Cambridge handbook of multimedia learning* 2.1 (2005), p. 24.

[219] Kathryn McAlindon et al. "The bond framework: a practical application of visual communication design and marketing to advance evaluation reporting". In: *American Journal of Evaluation* 40.2 (2019), pp. 291–305.

[220] George A. McCulley. "Writing Quality, Coherence, and Cohesion". In: *Research in the Teaching of English* 19.3 (1985). Publisher: National Council of Teachers of English, pp. 269–282. ISSN: 0034-527X. URL: https://www.jstor.org/stable/40171050 (visited on 09/14/2023).

[221] Deborah McCutchen, Paul Teske, and Catherine Bankston. "Writing and cognition: Implications of the cognitive architecture for learning to write and writing to learn". In: *Handbook of research on writing*. Routledge, 2009, pp. 554–578.

[222] Mary L. McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia Medica* 22.3 (2012), pp. 276–282.

[223] John McPhee. *Draft No. 4: On the writing process*. Farrar, Straus and Giroux, 2017.

[224] Bertalan Meskó and Eric J Topol. "The imperative for regulatory oversight of large language models (or generative AI) in healthcare". In: *NPJ digital medicine* 6.1 (2023), p. 120.

[225] Piotr Mirowski et al. "Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581225. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544548.3581225.

[226] Mohamed Zulhilmi bin Mohamed et al. "Artificial intelligence in mathematics education: A systematic literature review". In: *International Electronic Journal of Mathematics Education* 17.3 (2022), em0694.

[227] Ethan Mollick and Ethan Mollick. *Co-Intelligence*. Random House UK, 2024.

[228] Steven Moore et al. "Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods". In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 229–245.

[229] Caterina Moruzzi and Solange Margarido. "A User-centered Framework for Human-AI Co-creativity". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3650929. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3650929.

[230] Benjamin A Motz et al. "Terracotta: A tool for conducting experimental research on student learning". In: *Behavior Research Methods* 56.3 (2024), pp. 2519–2536.

[231] Junaid Mubeen. *Mathematical Intelligence: What We Have that Machines Don't*. Profile Books, 2022.

[232] Michael Muller and Justin Weisz. "Extending a Human-AI Collaboration Framework with Dynamism and Sociality". In: *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*. CHIWORK '22. Durham, NH, USA: Association for Computing Machinery, 2022. ISBN: 9781450396554. DOI: 10.1145/3533406.3533407. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3533406.3533407.

[233] April Murphy, Stephen E Fancsali, and Steven Ritter. "UpGrade: Scalable Digital Experimentation in Authentic Learning Settings". In: ().

[234] Ilya Musabirov et al. "Platform-based Adaptive Experimental Research in Education: Lessons Learned from The Digital Learning Challenge". In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 2025, pp. 13–23.

[235] National Council of Teachers of Mathematics. *Artificial Intelligence and Mathematics Teaching*. Available at https://www.nctm.org/standards-and-positions/Position-Statements/Artificial-Intelligence-and-Mathematics-Teaching/. 2024.

[236] Seyed Parsa Neshaei et al. "Enhancing Peer Review with AI-Powered Suggestion Generation Assistance: Investigating the Design Dynamics". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. IUI '24. Greenville, SC, USA: Association for Computing Machinery, 2024, pp. 88–102. ISBN: 9798400705083. DOI: 10.1145/3640543.3645169. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3640543.3645169.

[237] Philip M Newton and Maira Xiromeriti. "ChatGPT performance on MCQ exams in higher education. A pragmatic scoping review". In: *EdArXiv. February* 21 (2023).

[238] Andres Neyem et al. "Exploring the Impact of Generative AI for StandUp Report Recommendations in Software Capstone Project Development". In: *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. SIGCSE 2024. Portland, OR, USA: Association for Computing Machinery, 2024, pp. 951–957. ISBN: 9798400704239. DOI: 10.1145/3626252.3630854. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3626252.3630854.

[239] Huy A Nguyen et al. "Evaluating chatgpt's decimal skills and feedback generation in a digital learning game". In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 278–293.

[240] Eric Nichols, Leo Gao, and Randy Gomez. "Collaborative Storytelling with Large-scale Neural Language Models". In: *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*. MIG '20. Virtual Event, SC, USA: Association for Computing Machinery, 2020. ISBN: 9781450381710. DOI: 10.1145/3424636.3426903. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3424636.3426903.

[241] Jakob Nielsen and Jan Maurits Faber. "Improving system usability through parallel design". In: *Computer* 29.2 (1996), pp. 29–35.

[242] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[243] Zahra Nouri et al. "Supporting Requesters in Writing Clear Crowdsourcing Task Descriptions Through Computational Flaw Assessment". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. Sydney, NSW, Australia: Association for Computing Machinery, 2023, pp. 737–749. ISBN: 9798400701061. DOI: 10.1145/3581641.3584039. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3581641.3584039.

[244] MARTIN NYSTRAND. "A Social-Interactive Model of Writing". In: *Written Communication* 6.1 (1989), pp. 66–85. DOI: 10.1177/0741088389006001005. eprint: ttps://doi.org/10.1177/0741088389006001005. URL: https://doi.org/10.1177/0741088389006001005.

[245] Kathleen O'Leary et al. ""Suddenly, We Got to Become Therapists for Each Other": Designing Peer Support Chats for Mental Health". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3173905. URL: https://doi.org/10.1145/3173574.3173905.

[246] Tenaha O'Reilly, Sonya Symons, and Heather MacLatchy-Gaudet. "A comparison of self-explanation and elaborative interrogation". In: *Contemporary Educational Psychology* 23.4 (1998), pp. 434–445.

[247] OATutor. *OATutor: Open Adaptive Tutoring for Teachers*. Available at https://www.oatutor.io/teachers (accessed December 10, 2024). 2024.

[248] OpenAI. URL: https://platform.openai.com/docs/guides/prompt-engineering/strategy-split-complex-tasks-into-simpler-subtasks.

[249] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[250] OpenAI. "Introducing Structured Outputs in the API". In: *OpenAI* (Aug. 2024). URL: https://openai.com/index/introducing-structured-outputs-in-the-api/.

[251] OpenStax. *OpenStax College Algebra 2e*. URL: https://academia.stackexchange.com/questions/14010/how-do-you-cite-a-github-repository.

[252] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. "BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380959. DOI: 10.1145/3411763.3450391. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3411763.3450391.

[253] Matthew J Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n71. eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf. URL: https://www.bmj.com/content/372/bmj.n71.

[254] Philip Pallmann et al. "Adaptive designs in clinical trials: why use them, and how to run and report them". In: *BMC medicine* 16.1 (2018), pp. 1–15.

[255] *PaLM 2 Technical Report*. https://ai.google/static/documents/palm2techreport.pdf. May 2023.

[256] John F Pane et al. "Effectiveness of cognitive tutor algebra I at scale". In: *Educational Evaluation and Policy Analysis* 36.2 (2014), pp. 127–144.

[257] Gilbert Paquette et al. "Competency-based personalization for massive online learning". In: *Smart Learning Environments* 2.1 (2015), pp. 1–19.

[258] Zachary A Pardos and Shreya Bhandari. "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills". In: *Plos one* 19.5 (2024), e0304013.

[259] Zachary A Pardos and Shreya Bhandari. "Learning gain differences between ChatGPT and human tutor generated algebra hints". In: *arXiv preprint arXiv:2302.06871* (2023).

[260] Zachary A Pardos et al. "Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework". In: *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 2017, pp. 23–32.

[261] Zachary A Pardos et al. "Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research". In: *Proceedings of the 2023 chi conference on human factors in computing systems*. Association for Computing Machinery. 2023, pp. 1–17. DOI: https://doi.org/10.1145/3544548.3581574.

[262] Zachary A. Pardos. *Open-source Adaptive Tutoring System (OATutor)*. https://github.com/CAHLR/OATutor-Content/tree/. 2023.

[263] Hyanghee Park and Daehwan Ahn. "The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642785. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642785.

[264] Gigliola Paviotti, Pier Giuseppe Rossi, and Dénes Zarka. "Intelligent tutoring systems: an overview". In: *Pensa Multimedia* (2012).

[265] Zhenhui Peng et al. "Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–15. ISBN: 9781450367080. DOI: 10.1145/3313831.3376695. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3313831.3376695.

[266] Zhenhui Peng et al. "Storyfier: Exploring Vocabulary Learning Support with Text Generation Models". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST '23. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606786. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3586183.3606786.

[267] Penzance. *A/B Tool GitHub Repository*. https://github.com/penzance/ab-testing-tool/wiki/Getting-Started. Last accessed: 22-02-2021. 2021.

[268] Juanan Pereira and Mikel Alejo Barcina. "A chatbot assistant for writing good quality technical reports". In: *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM'19. León, Spain: Association for Computing Machinery, 2019, pp. 59–64. ISBN: 9781450371919. DOI: 10.1145/3362789.3362798. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3362789.3362798.

[269] Andrew Petersen. *Programming Course Resource System (PCRS)*. Accessed: 2025-06-09. 2013. URL: https://mcs.utm.utoronto.ca/~pcrs/pcrs/.

[270] Ritika Poddar et al. "AI Writing Assistants Influence Topic Choice in Self-Presentation". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3585893. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3544549.3585893.

[271] James P Purdy. "What can design thinking offer writing studies?" In: *College Composition and Communication* (2014), pp. 612–641.

[272] Hua Xuan Qin et al. "CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642105. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642105.

[273] Muhammad Raees et al. "From explainable to interactive AI: A literature review on current trends in human-AI interaction". In: *International Journal of Human-Computer Studies* (2024), p. 103301.

[274] Anna Rafferty, Huiji Ying, and Joseph Williams. "Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments". In: *JEDM— Journal of Educational Data Mining* 11.1 (2019), pp. 47–79.

[275] Naveen Ram et al. "Say What? Collaborative Pop Lyric Generation Using Multitask Transfer Learning". In: *Proceedings of the 9th International Conference on Human-Agent Interaction*. HAI '21. Virtual Event, Japan: Association for Computing Machinery, 2021, pp. 165–173. ISBN: 9781450386203. DOI: 10.1145/3472307.3484175. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3472307.3484175.

[276] Sarah Ransdell, C. Michael Levy, and Ronald T. Kellogg. "The structure of writing processes as revealed by secondary task demands". en. In: *L1-Educational Studies in Language and Literature* 2.2 (May 2002), pp. 141–163. ISSN: 1573-1731. DOI: 10.1023/A:1020851300668. URL: https://doi.org/10.1023/A:1020851300668 (visited on 09/14/2023).

[277] Christian Rapp et al. "Thesis Writer: A System for Supporting Academic Writing". In: *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. 2015, pp. 57–60.

[278] Swati Rawat et al. "Exploring the Potential of ChatGPT to improve experiential learning in Education". In: *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*. ICIMMI '23. Jaipur, India: Association for Computing Machinery, 2024. ISBN: 9798400709418. DOI: 10.1145/3647444.3647910. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3647444.3647910.

[279] Leena Razzaq et al. "The Assistment Builder: Supporting the life cycle of tutoring system content creation". In: *IEEE Transactions on Learning Technologies* 2.2 (2009), pp. 157–166.

[280] *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/. Feb. 2013.

[281] Olaf Resch and Aglika Yankova. "Open knowledge interface: a digital assistant to support students in writing academic assignments". In: *Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence*. EASEAI 2019. Tallinn, Estonia: Association for Computing Machinery, 2019, pp. 13–16. ISBN: 9781450368520. DOI: 10.1145/3340435.3342723. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3340435.3342723.

[282] Mohi Reza et al. "ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–18.

[283] Mohi Reza et al. "Experimenting with Experimentation: Rethinking The Role of Experimentation in Educational Design". In: *arXiv preprint arXiv:2208.05069* (2022).

[284] Mohi Reza et al. "The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses". In: *Proceedings of the Eighth ACM Conference on Learning @ Scale*. L@S '21. Virtual Event, Germany: Association for Computing Machinery, 2021, pp. 15–26. ISBN: 9781450382151. DOI: 10.1145/3430895.3460128. URL: https://doi.org/10.1145/3430895.3460128.

[285] Jeba Rezwana and Mary Lou Maher. "User Perspectives on Ethical Challenges in Human-AI Co-Creativity: A Design Fiction Study". In: *Proceedings of the 15th Conference on Creativity and Cognition*. C&C '23. Virtual Event, USA: Association for Computing Machinery, 2023, pp. 62–74. ISBN: 9798400701801. DOI: 10.1145/3591196.3593364. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3591196.3593364.

[286] Horst Rittel. "Wicked problems". In: *Management Science,(December 1967)* 4.14 (1967).

[287] Steven Ritter et al. "UpGrade: An open source tool to support A/B testing in educational software". In: *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*. 2020.

[288] Ronald E Robertson et al. ""I Can't Reply with That": Characterizing Problematic Email Reply Suggestions". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445557. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3411764.3445557.

[289] David Robotham and Claire Julian. "Stress and the higher education student: a critical review of the literature". In: *Journal of further and higher education* 30.02 (2006), pp. 107–117.

[290] Melissa Roderick. "Drowning in data but thirsty for analysis". In: *Teachers College Record* 114.11 (2012), pp. 1–9.

[291] Melissa Roemmele and Andrew Gordon. "Creative Help: A Story Writing Assistant". In: Nov. 2015, pp. 81–92. ISBN: 978-3-319-27035-7. DOI: 10.1007/978-3-319-27036-4_8.

[292] D. Gordon Rohman. "Pre-Writing the Stage of Discovery in the Writing Process". In: *College Composition and Communication* 16.2 (1965), pp. 106–112. ISSN: 0010096X. URL: http://www.jstor.org/stable/354885 (visited on 10/29/2024).

[293] Abel Salinas and Fred Morstatter. "The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance". In: *arXiv preprint arXiv:2401.03729* (2024).

[294] Nithya Sambasivan and Rajesh Veeraraghavan. "The deskilling of domain expertise in AI development". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–14.

[295] Bahareh Sarrafzadeh et al. "Characterizing Stage-aware Writing Assistance for Collaborative Document Authoring". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW3 (Jan. 2021). DOI: 10.1145/3434180. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3434180.

[296] Dirk Schmidt. "Grading Tibetan Children's Literature: A Test Case Using the NLP Readability Tool "Dakje"". In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19.6 (Oct. 2020). ISSN: 2375-4699. DOI: 10.1145/3392046. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3392046.

[297] Oliver Schmitt and Daniel Buschek. "CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot". In: *Proceedings of the 13th Conference on Creativity and Cognition*. C&C '21. Virtual Event, Italy: Association for Computing Machinery, 2021. ISBN: 9781450383769. DOI: 10.1145/3450741.3465253. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3450741.3465253.

[298] Good Science. *Why We Still Need Many More RCTs*. Accessed: 2024-10-05. 2024. URL: https://goodscience.substack.com/p/why-we-still-need-many-more-rcts.

[299] Emily E Scott, Mary Pat Wenderoth, and Jennifer H Doherty. "Design-based research: A methodology to extend and enrich biology education research". In: *CBE—Life Sciences Education* 19.2 (2020), es11.

[300] Steven L. Scott. "Multi-armed bandit experiments in the online service economy". In: *Applied Stochastic Models in Business and Industry* 31 (2015). Special issue on actual impact and future perspectives on stochastic modelling in business and industry, pp. 37–49. URL: http://onlinelibrary.wiley.com/doi/10.1002/asmb.2104/abstract.

[301] Anthony Seow. "The writing process and process writing". In: *Methodology in language teaching: An anthology of current practice* 315 (2002), p. 320.

[302] Orit Shaer et al. "AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–17.

[303] Priten Shah. *AI and the Future of Education: Teaching in the Age of Artificial Intelligence*. John Wiley & Sons, 2023.

[304] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. "SAGA: Collaborative Storytelling with GPT-3". In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '21 Companion. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 163–166. ISBN: 9781450384797. DOI: 10.1145/3462204.3481771. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3462204.3481771.

[305] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. "Saga: Collaborative storytelling with gpt-3". In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 2021, pp. 163–166.

[306] Mike Sharples. *How we write: Writing as creative design*. Routledge, 2002.

[307] Shreya Sheel, Ioannis Anastasopoulos, and Zach A Pardos. "Comparing Authoring Experiences with Spreadsheet Interfaces vs GUIs". In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 2024, pp. 598–607.

[308] Hua Shen et al. "ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing". In: *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '23 Companion. Minneapolis, MN, USA: Association for Computing Machinery, 2023, pp. 384–387. ISBN: 9798400701290. DOI: 10.1145/3584931.3607492. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3584931.3607492.

[309] Thomas B Sheridan, William L Verplank, and TL Brooks. "Human/computer control of undersea teleoperators". In: *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*. 1978.

[310] Thomas B Sheridan, William L Verplank, and TL Brooks. "Human/computer control of undersea teleoperators". In: *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*. 1978.

[311] Antonette Shibani et al. "Untangling Critical Interaction with AI in Students' Written Assessment". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: 10.1145/3613905.3651083. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613905.3651083.

[312] Donghoon Shin et al. "Exploring the Effects of AI-assisted Emotional Support Processes in Online Mental Health Community". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391566. DOI: 10.1145/3491101.3519854. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491101.3519854.

[313] Hyungyu Shin et al. "Understanding the effect of in-video prompting on learners and instructors". In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–12.

[314] Ben Shneiderman. "Human-centered AI: ensuring human control while increasing automation". In: *Proceedings of the 5th Workshop on Human Factors in Hypertext*. HUMAN '22. Barcelona, Spain: Association for Computing Machinery, 2022. ISBN: 9781450394017. DOI: 10.1145/3538882.3542790. URL: https://doi.org/10.1145/3538882.3542790.

[315] Ben Shneiderman. "Human-centered AI: ensuring human control while increasing automation". In: *Proceedings of the 5th Workshop on Human Factors in Hypertext*. 2022, pp. 1–2.

[316] Ben Shneiderman. "Human-centered artificial intelligence: Reliable, safe & trustworthy". In: *International Journal of Human–Computer Interaction* 36.6 (2020), pp. 495–504.

[317] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. "Can LLMs Generate Novel Research Ideas?" In: *arXiv preprint arXiv:2409.04109* (2024).

[318] Herbert A Simon. "What we know about learning". In: *Journal of Engineering Education* 87.4 (1998), pp. 343–348.

[319] Anjali Singh et al. "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. Kyoto, Japan: Association for Computing Machinery, 2024, pp. 742–748. ISBN: 9798400716188. DOI: 10.1145/3636555.3636853. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3636555.3636853.

[320] Joykirat Singh, Akshay Nambi, and Vibhav Vineet. "Exposing the Achilles' Heel: Evaluating LLMs Ability to Handle Mistakes in Mathematical Reasoning". In: *arXiv preprint arXiv:2406.10834* (2024).

[321] Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. "FigurA11y: AI Assistance for Writing Scientific Alt Text". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. IUI '24. Greenville, SC, USA: Association for Computing Machinery, 2024, pp. 886–906. ISBN: 9798400705083. DOI: 10.1145/3640543.3645212. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3640543.3645212.

[322] Nikhil Singh et al. "Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence". In: *ACM Trans. Comput.-Hum. Interact.* 30.5 (Sept. 2023). ISSN: 1073-0516. DOI: 10.1145/3511599. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3511599.

[323] Dan Siroker et al. *Systems and methods for website optimization*. US Patent 8,839,093. Sept. 2014.

[324] Gary Smith. *The AI delusion*. Oxford University Press, 2018.

[325] Nancy Sommers. "Revision strategies of student writers and experienced adult writers". In: *College composition and communication* 31.4 (1980), pp. 378–388.

[326] Changwon Son et al. "Effects of COVID-19 on college students' mental health in the United States: Interview survey study". In: *Journal of medical internet research* 22.9 (2020), e21279.

[327] Daniel Stafford and Robert Flatley. "Openstax". In: *The Charleston Advisor* 20.1 (2018), pp. 48–51.

[328] John Stamper et al. "Experiments as a service Infrastructure (EASI)." In: (Sept. 2023). DOI: 10.1184/R1/24205188.v1. URL: https://kilthub.cmu.edu/articles/poster/Experiments_as_a_service_Infrastructure_EASI_/24205188.

[329] John Stamper et al. "LearnSphere: A Learning Data and Analytics Cyberinfrastructure". In: *Journal of Educational Data Mining* 16.1 (2024), pp. 141–163.

[330] John C Stamper et al. "Managing the educational dataset lifecycle with DataShop". In: *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*. Springer. 2011, pp. 557–559.

[331] Joachim Stoeber and Dirk P Janssen. "Perfectionism and coping with daily failures: Positive reframing helps achieve satisfaction at the end of the day". In: *Anxiety, Stress & Coping* 24.5 (2011), pp. 477–497.

[332] Sangho Suh et al. "Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642400. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3642400.

[333] Lu Sun et al. "MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review". In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). DOI: 10.1145/3637371. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3637371.

[334] Reid Swanson and Andrew Gordon. "Say Anything". In: *ACM Transactions on Interactive Intelligent Systems* 2 (Sept. 2012), pp. 1–35. DOI: 10.1145/2362394.2362398.

[335] John Sweller, Paul A Kirschner, and Richard E Clark. "Why minimally guided teaching techniques do not work: A reply to commentaries". In: *Educational psychologist* 42.2 (2007), pp. 115–121.

[336] Mohammed Taiye et al. "Generative AI-Enhanced Academic Writing: A Stakeholder-Centric Approach for the Design and Development of CHAT4ISP-AI". In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC '24. Avila, Spain: Association for Computing Machinery, 2024, pp. 74–80. ISBN: 9798400702433. DOI: 10.1145/3605098.3636055. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3605098.3636055.

[337] Yuying Tang et al. "Exploring the Impact of AI-generated Image Tools on Professional and Non-professional Users in the Art and Design Fields". In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. CSCW Companion '24. San Jose, Costa Rica: Association for Computing Machinery, 2024, pp. 451–458. ISBN: 9798400711145. DOI: 10.1145/3678884.3681890. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3678884.3681890.

[338] Lev Tankelevitch et al. "The Metacognitive Demands and Opportunities of Generative AI". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642902. URL: https://doi.org/10.1145/3613904.3642902.

[339] Cem Tekin, Jonas Braun, and Mihaela van der Schaar. "etutor: Online learning for personalized education". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 5545–5549.

[340] Khanh-Phuong Thai et al. "Learning engineering is human-centered". In: *Learning engineering toolkit*. Routledge, 2022, pp. 83–123.

[341] The Learning Agency. *Introduction to Learning Engineering*. Guide & Resource. Accessed: 2025-06-08. 2025. URL: https://the-learning-agency.com/guides-resources/introduction-to-learning-engineering/.

[342] Jakob Tholander and Martin Jonsson. "Design Ideation with AI - Sketching, Thinking and Talking with Generative Machine Learning Models". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. DIS '23. Pittsburgh, PA, USA: Association for Computing Machinery, 2023, pp. 1930–1940. ISBN: 9781450398930. DOI: 10.1145/3563657.3596014. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3563657.3596014.

[343] William R Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3-4 (1933), pp. 285–294.

[344] Karran Thorpe. "Reflective learning journals: From concept to practice". In: *Reflective practice* 5.3 (2004), pp. 327–343.

[345] Maryam Tohidi et al. "Getting the Right Design and the Design Right". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 1243–1252. ISBN: 1595933727. DOI: 10.1145/1124772.1124960. URL: https://doi.org/10.1145/1124772.1124960.

[346] Maryam Tohidi et al. "Getting the right design and the design right". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 1243–1252. ISBN: 1595933727. DOI: 10.1145/1124772.1124960. URL: https://doi.org/10.1145/1124772.1124960.

[347] Hugo Touvron, Louis Martin, and Kevin Stone. "Llama 2: Open Foundation and Fine-Tuned Chat Models". en. In: ().

[348] Allison S Troy et al. "Cognitive reappraisal and acceptance: Effects on emotion, physiology, and perceived cognitive costs." In: *Emotion* 18.1 (2018), p. 58.

[349] Edward R Tufte. "Envisioning Information Graphics Press". In: *Cheshire, Connecticut* 6410 (1990), pp. 1–35.

[350] United Nations. *Digital Learning for All: Transforming Education*. Accessed: 2024-09-28. 2024. URL: https://www.un.org/en/transforming-education-summit/digital-learning-all.

[351] Konstantinos Vassakis, Emmanuel Petrakis, and Ioannis Kopanakis. "Big data analytics: applications, prospects and challenges". In: *Mobile big data*. Springer, 2018, pp. 3–20.

[352] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[353] Joannes Vermorel and Mehryar Mohri. "Multi-armed bandit algorithms and empirical evaluation". In: *European conference on machine learning*. Springer. 2005, pp. 437–448.

[354] Gregory M Walton and Timothy D Wilson. "Wise interventions: Psychological remedies for social and personal problems." In: *Psychological review* 125.5 (2018), p. 617.

[355] Thiemo Wambsganss et al. "Adaptive Empathy Learning Support in Peer Review Scenarios". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517740. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3517740.

[356] Thiemo Wambsganss et al. "AL: An Adaptive Learning Support System for Argumentation Skills". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376732. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3313831.3376732.

[357] Thiemo Wambsganss et al. "ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445781. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3411764.3445781.

[358] Qian Wan et al. ""It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models". In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). DOI: 10.1145/3637361. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3637361.

[359] Jiyao Wang et al. "Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3641917. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3613904.3641917.

[360] Ruochen Wang et al. "One Prompt is not Enough: Automated Construction of a Mixture-of-Expert Prompts". In: *arXiv preprint arXiv:2407.00256* (2024).

[361] Xuezhi Wang et al. "Self-consistency improves chain of thought reasoning in language models". In: *arXiv preprint arXiv:2203.11171* (2022).

[362] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. "Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517551. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3491102.3517551.

[363] Zijie J Wang et al. "Wordflow: Social prompt engineering for large language models". In: *arXiv preprint arXiv:2401.14447* (2024).

[364] Azmine Toushik Wasi, Taki Hasan Rafi, and Dong-Kyu Chae. "DiaFrame: A Framework for Understanding Bengali Dialects in Human-AI Collaborative Creative Writing Spaces". In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. CSCW Companion '24. San Jose, Costa Rica: Association for Computing Machinery, 2024, pp. 268–274. ISBN: 9798400711145. DOI: 10.1145/3678884.3681862. URL: https://doi.org/10.1145/3678884.3681862.

[365] Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. "An interaction design for machine teaching to develop AI tutors". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–11.

[366] David Wiley, TJ Bliss, and Mary McEwen. "Open educational resources: A review of the literature". In: *Handbook of research on educational communications and technology* (2013), pp. 781–789.

[367] Joseph Jay Williams et al. "Axis: Generating explanations at scale with learnersourcing and machine learning". In: *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. 2016, pp. 379–388.

[368] Joseph Jay Williams et al. "Enhancing online problems through instructor-centered tools for randomized experiments". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–12.

[369] Joost FC de Winter and Dimitra Dodou. "Five-point likert items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012)". In: *Practical Assessment, Research, and Evaluation* 15.1 (2010), p. 11.

[370] Conrad Wolfram. *The math (s) fix: An education blueprint for the AI age*. Wolfram Media, 2020.

[371] Shaomei Wu et al. "Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. DOI: 10.1145/3290605.3300746. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3290605.3300746.

[372] Yi Wu et al. "How Effective Are Neural Networks for Fixing Security Vulnerabilities". In: *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. ISSTA 2023. Seattle, WA, USA: Association for Computing Machinery, 2023, pp. 1282–1294. ISBN: 9798400702211. DOI: 10.1145/3597926.3598135. URL: https://doi.org/10.1145/3597926.3598135.

[373] Yiying Wu, Yunye Yu, and Pengcheng An. "Dancing with the Unexpected and Beyond: The Use of AI Assistance in Design Fiction Creation". In: *Proceedings of the Tenth International Symposium of Chinese CHI*. Chinese CHI '22. Guangzhou, China and Online, China: Association for Computing Machinery, 2024, pp. 129–140. ISBN: 9781450398695. DOI: 10.1145/3565698.3565777. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3565698.3565777.

[374] xAPI. *xAPI*. Last accessed: 22-02-2021. 2021. URL: https://xapi.com/overview/.

[375] XPRIZE. *Digital Learning Challenge*. Accessed: 2024-11-15. 2024. URL: https://www.xprize.org/challenge/digitallearning.

[376] Xiaotong (Tone) Xu et al. "Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. IUI '24. Greenville, SC, USA: Association for Computing Machinery, 2024, pp. 907–921. ISBN: 9798400705083. DOI: 10.1145/3640543.3645196. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3640543.3645196.

[377] Pinar Yanardag, Manuel Cebrian, and Iyad Rahwan. "Shelley: A Crowd-sourced Collaborative Horror Writer". In: *Proceedings of the 13th Conference on Creativity and Cognition*. C&C '21. Virtual Event, Italy: Association for Computing Machinery, 2021. ISBN: 9781450383769. DOI: 10.1145/3450741.3465251. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3450741.3465251.

[378] Ann Yuan et al. "Wordcraft: Story Writing With Large Language Models". In: *27th International Conference on Intelligent User Interfaces*. IUI '22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 841–852. ISBN: 9781450391443. DOI: 10.1145/3490099.3511105. URL: https://doi.org/10.1145/3490099.3511105.

[379] Ann Yuan et al. "Wordcraft: Story Writing With Large Language Models". In: *Proceedings of the 27th International Conference on Intelligent User Interfaces*. IUI '22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 841–852. ISBN: 9781450391443. DOI: 10.1145/3490099.3511105. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3490099.3511105.

[380] J.D. Zamfirescu-Pereira et al. "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581388. URL: https://doi.org/10.1145/3544548.3581388.

[381] J.D. Zamfirescu-Pereira et al. "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581388. URL: https://doi.org/10.1145/3544548.3581388.

[382] Angela Zavaleta-Bernuy et al. "Using adaptive experiments to rapidly help students". In: *International Conference on Artificial Intelligence in Education*. Springer. 2021, pp. 422–426.

[383] Mingyuan Zhang et al. "Towards Human-Centred AI-Co-Creation: A Three-Level Framework for Effective Collaboration between Human and AI". In: *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '23 Companion. Minneapolis, MN, USA: Association for Computing Machinery, 2023, pp. 312–316. ISBN: 9798400701290. DOI: 10.1145/3584931.3607008. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3584931.3607008.

[384] Peng Zhang and Maged N Kamel Boulos. "Generative AI in medicine and healthcare: promises, opportunities and challenges". In: *Future Internet* 15.9 (2023), p. 286.

[385] Zheng Zhang et al. "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. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606800. URL: https://doi-org.myaccess.library.utoronto.ca/10.1145/3586183.3606800.

[386] Eric Zhou and Dokyun Lee. "Generative artificial intelligence, human creativity, and art". In: *PNAS nexus* 3.3 (2024), pgae052.

[387] William Zinsser. "On writing well: The classic guide to writing nonfiction". In: *New York, NY* (2006).

[388] Jesse S Zolna. "The effects of study time and presentation modality on learning". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA. Los Angeles, CA: SAGE Publications Sage CA, 2007, pp. 512–515.