석 사 학 위 논 문
Master's Thesis

대형 언어 모델 보조 글쓰기에서 저자성 변화와 자기효능감·신뢰의 발전 과정 이해

Authorship Drift: How Self-Efficacy and Trust Evolve During LLM-Assisted Writing

2026

박 연 수 (朴 延 洙 Park, Yeon Su)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

대형 언어 모델 보조 글쓰기에서 저자성 변화와 자기효능감·신뢰의 발전 과정 이해

2026

박 연 수

한 국 과 학 기 술 원

전산학부

대형 언어 모델 보조 글쓰기에서 저자성 변화와 자기효능감·신뢰의 발전 과정 이해

박 연 수

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2025년 12월 9일

심사위원장 김 주 호 (인)

심 사 위 원 오 혜 연 (인)

심 사 위 원 이 의 진 (인)

Authorship Drift: How Self-Efficacy and Trust Evolve During LLM-Assisted Writing

Yeon Su Park

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 9, 2025

Approved by

Juho Kim
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS

박연수. 대형 언어 모델 보조 글쓰기에서 저자성 변화와 자기효능감·신뢰의 발전 과정 이해. 전산학부 . 2026년. 37+iv 쪽. 지도교수: 김주호. (영문 논문)

Yeon Su Park. Authorship Drift: How Self-Efficacy and Trust Evolve During LLM-Assisted Writing. School of Computing . 2026. 37+iv pages. Advisor: Juho Kim. (Text in English)

초 록

대형 언어 모델(LLM)은 주관적 글쓰기에서 핵심적인 협업 파트너로 자리 잡아가고 있다. 이러한 순간순간의 협업 결정—모델의 능력을 활용할 것인지, 혹은 글의 저자성을 지키기 위해 스스로 작성할 것인지—은 사용자의 내적 상태, 특히 자기효능감(Self-Efficacy)과 모델에 대한 신뢰(Trust)의 변화를 수반한다. 그러나 이 두 상태가 시간에 따라 어떻게 변화하고, 이 변화가 협업 과정에서의 전략적 결정과 어떻게 연결되는지는 충분히 탐구되지 않았다. 본 연구에서는 302명을 대상으로 LLM 기반 글쓰기 과제를 수행하게 하고, 상호작용 로그와 턴 단위 자기효능감·신뢰 점수를 수집하여 이러한 역동적 변화를 분석했다. 분석 결과, LLM과의 협업은 전반적으로 사용자들의 자기효능감을 감소시키는 반면 신뢰는 증가시키는 경향이 있었다. 자기효능감이 감소한 참가자들은 자신의 글을 모델이 직접 수정해주도록 요청할 가능성이 더 높았으며, 반대로 자기효능감을 회복한 참가자들은 리뷰나 피드백을 더 많이 요청했다. 또한, 자기효능감이 안정적으로 유지된 참가자들은 최종 결과물에서 훨씬 더 높은 수준의 저자성을 보고했다. 본 연구는 이러한 결과를 바탕으로 인간–LLM 협업 맥락에서 사용자의 저자성과 주도성을 이해하고 지원하기 위한 시사점을 제안한다.

핵 심 낱 말 자기효능감, 신뢰, 저자성, LLM 상호작용 패턴, LLM 기반 글쓰기

Abstract

Large language models (LLMs) have become integral collaborators in open-ended writing tasks. These moment-to-moment collaborative decisions—between leveraging the model's capabilities and preserving authorship of their writing—involve fluctuations in users' internal states, particularly in self-efficacy and trust in the model. However, the dynamics of these states and their association with strategic decisions during collaboration remain underexplored. We examine these dynamics in a study with 302 participants on an LLM-assisted writing task, logging interactions and per-turn self-efficacy and trust ratings. Our analysis shows that collaboration generally decreased users' self-efficacy while increasing trust. Participants who lost self-efficacy were more likely to ask the LLM to edit their work directly, whereas those who recovered self-efficacy requested more review and feedback. Furthermore, participants whose self-efficacy remained stable reported substantially higher authorship of the final result. Based on these findings, we propose design implications for understanding and supporting authorship in human-LLM collaboration.

Keywords Self-Efficacy, Trust, Authorship, LLM Interaction Pattern, LLM-Assisted Writing

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Large language models (LLMs) have emerged as powerful general-purpose tools, leveraging their natural language capabilities and vast knowledge across diverse domains [1, 2]. Hence, LLMs are now perceived as collaborative partners who are co-producing the content with the users [3], extending their individual capabilities.

This collaborative potential is particularly visible in writing, a domain where ideas, structure, and phrasing are continuously negotiated and refined, making it a cognitively demanding and inherently iterative activity [4, 5]. Given this complexity, LLMs can play a complementary role in easing the writer's mental load and assisting performance. Many previous studies highlighted the benefits of such human-LLM collaborative writing, including enhanced productivity and writer confidence [6], improved writing quality [7], and facilitated idea exploration [8].

However, LLM-assisted writing introduces critical challenge of authorship. Here, users and LLMs continuously negotiate the content across multiple turns. As LLMs take on more generative and revisional roles [9, 10], boundaries between user and model contributions become less clear. In subjective tasks like writing where multiple solutions can be equally plausible, users may find it difficult to critically evaluate or distinguish the model's input from their own. This blurring can lead to uncritical adoption of AI-generated ideas or phrasing [11], as well as a decrease in the sense of agency and ownership of the final result [12, 13]. This can further leave writers feeling disconnected from their work [12] and raise concerns about the originality and integrity of the content they produce [14].

Understanding this authorship problem requires attending to the writer's evolving internal state during collaboration. Recent work on human–LLM collaboration highlights two complementary psychological constructs in this process: *self-efficacy*—users' belief in their ability to accomplish the work on their own—and *trust*—their belief that the LLM will reliably support them [15, 16]. However, most existing studies treat self-efficacy and trust as static attributes, overlooking the fact that writing with an LLM is a dynamic process. As writers negotiate ideas and revisions across multiple conversational turns, their perceptions of both themselves and the model can shift substantially [17, 9, 11]. Without capturing these moment-to-moment changes and how they evolve over time, we cannot fully explain how authorship weakens, strengthens, or transforms throughout the collaborative writing process.

To this end, we conducted an online experiment with 302 participants examining the turn-by-turn dynamics of self-efficacy and trust in LLM-assisted writing. While self-efficacy and trust are traditionally conceptualized as stable psychological traits [18, 19], we adopt a task-specific interpretation to capture these constructs as dynamic states that fluctuate across turns. In our study, we define *self-efficacy* as the writer's momentary belief in their ability to complete this task on their own, and *trust* as their momentary belief that the LLM will reliably support this task. This task-specific view enables us to trace the dynamic trajectories of self-efficacy and trust throughout the writing process. Through this lens, we identify distinct patterns in how these constructs evolve, and examine how these patterns relate to users' prompting strategies and to both actual and perceived authorship of their produced text.

Our findings reveal distinct trajectory patterns: while self-efficacy tends to decrease over interaction turns, trust generally increases. We found that users with different self-efficacy trajectories exhibit distinct behavioral patterns. Participants with declining self-efficacy were more likely to ask the LLM for direct edits, often employing draft-to-edit and edit-to-edit prompt sequences, which resulted in lower authorship. Participants who experienced initial self-efficacy drops but recovered were more likely to request review and feedback from the LLM. In contrast, participants with stable self-efficacy reported substantially higher authorship of the final result overall. Based on these insights, we outline design implications for LLM-assisted writing systems that draw on temporal self-efficacy trajectories and behavioral signals to identify moments when authorship may be vulnerable and support users in maintaining agency.

The contributions of this paper are as follows:

- Empirical characterization of turn-level trajectories of self-efficacy and trust in LLM-assisted writing.
- Findings on relationship between self-efficacy trajectories and (1) users' prompting strategies, and (2) their actual and perceived authorship.
- Design implications for understanding and supporting authorship in human-LLM collaboration.

Chapter 2. Related Works

We review the foundational literature across three domains that inform and contextualize our investigation of dynamic human-AI collaboration in writing tasks: (1) self-efficacy and trust in human-AI interactions, (2) usage strategies in collaborative AI tools, and (3) measuring authorship in collaborative writing with AI.

2.1 Self-Efficacy and Trust in Human-AI Collaboration

When users collaborate with LLMs across multiple turns, their decisions about when to rely on the model versus themselves depend on perceptions of their own capabilities and the system's. To examine these underlying psychological processes, we ground our work in two foundational constructs: *self-efficacy* and *trust*.

Bandura [18, 20] defined self-efficacy as people's beliefs in their capability to organize and execute the actions required to accomplish a task. This construct has been shown to shape task choice, effort, and persistence across domains such as education and work [21, 22]. In parallel, Mayer et al. [23] conceptualized trust as a willingness to be vulnerable to another agent based on the expectation that the agent will act in ways that matter to the trustor. Subsequent work in automation and HCI adopts this framing to describe how users decide whether to rely on AI systems under uncertainty [24, 25]. Together, self-efficacy and trust influence how people balance their own contributions with those of an AI system, shaping strategies for delegation, oversight, and joint performance [26, 15, 27].

A growing body of work in human–AI interaction has adopted these constructs, or closely related notions such as task-specific confidence, to explain how people engage with AI systems. Prior studies have examined how these beliefs shape selective adoption of AI suggestions [24], prompting strategies [28], reliance calibration [29, 16, 30], and decision-making behaviors [31, 32]. In most of this research, however, self-efficacy and trust are treated as static with respect to a given task. They are typically measured once or twice (e.g., pre–post) or inferred from behavior aggregated over the entire task, and are assumed to remain relatively stable throughout the interaction [33, 34]. As a result, moment-to-moment changes in these beliefs are rarely captured directly. While recent studies have begun to examine within-session adjustments in trust [35], work that jointly tracks how users' beliefs about themselves and the AI co-evolve during complex, multi-turn collaboration remains limited.

In this work, we adopt a process-oriented perspective and examine self-efficacy and trust as dynamic, task-specific states that evolve throughout interaction. By modeling their turn-level trajectories, we provide a fine-grained account of how beliefs about oneself and the AI shape collaborative behavior AI-assisted writing.

2.2 Usage Strategies in Collaborative AI Tools

AI-assisted collaboration transforms traditional linear workflows into iterative, non-sequential patterns where users move recursively between working stages [8, 36]. To understand what makes a workflow effective, prior research has primarily adopted an outcome-oriented approach, linking observable usage patterns to measurable results [37, 38]. For example, in education, measures often center on learning gains or exam performance [28, 39, 40, 41]. Creative tasks are often evaluated in terms of ideational fluency [42, 43], originality [44, 45], or diversity of output [43]. In decision-making contexts, success is typically measured through accuracy [24, 46, 16], efficiency [47, 31], or task completion rates [47].

However, as AI systems become increasingly capable at producing high-quality outputs [48, 49], outcome-oriented evaluation becomes less informative for understanding and improving human-AI interaction [50, 51]. When AI can reliably generate good results, the critical question shifts from 'what was produced' to 'how did the collaboration unfold' and 'what was the user's experience' [10]. In this context, examining user actions and their perceptions of those actions, such as a sense of authorship, becomes particularly important. Understanding the collaborative process becomes essential because it reveals whether users maintain meaningful agency [10] and cognitive engagement in the interaction [52], which directly impacts sustained learning [53] and long-term growth [54].

To understand these collaborative processes, researchers have begun examining user interactions through analysis of prompting patterns and conversational exchanges. For instance, McNichols et al. [55] have analyzed student-AI interactions to understand how different prompting strategies shape learning outcomes and engagement patterns. Similarly, Mysore et al. [9] analyze detailed traces of users' AI interactions and revisions to identify prompting styles and their relation to shifts in cognitive engagement and reliance. Building on this methodological approach, our study investigates how LLM prompting patterns relate to authorship as a process-oriented dimension, in order to understand which collaborative approaches foster meaningful human agency and sustained learning in AI-assisted writing.

2.3 Measuring Authorship in Collaborative Writing with AI

While previous research has shown that collaborative writing with AI improves authors' productivity and creative processes [44, 43, 8], central to this collaboration is the problem of authorship [13, 56, 10]. As AI systems function as coauthors, they blur the boundary between the contributions of humans and AI, potentially diminishing human agency [42] while complicating questions of ownership [13]. Hence, this raises fundamental questions about how to define, measure, and attribute authorship in human-AI collaborative writing contexts.

To address this, prior works have predominantly relied on perceptual approaches that capture users' subjective assessments of their authorship contributions. Studies typically employ surveys asking participants to rate their sense of ownership [12], creative input [57], or perceived contribution [13]. Interview-based methodologies have also been used to explore how writers negotiate boundaries between their work and AI-generated content [13, 12]. Some research has extended this approach by examining writers' attribution of agency and responsibility in collaborative outputs [56, 57].

However, these perceptual approaches have some limitations. Studies examining varying degrees of human versus AI contribution reveal differences between users' subjective authorship claims and their objective contribution levels [12, 56, 58], yet few studies have systematically examined both perceptual and objective dimensions together. In this work, we address this gap by jointly analyzing perceptual authorship and objective AI output utilization behaviors, establishing a more comprehensive measure of *actual* and *perceived* authorship in human–AI collaborative writing.

Chapter 3. Method

We investigate the dynamics of user self-efficacy and trust in the context of argumentative writing assisted by the LLM. As defined earlier (Section 1), we conceptualize *self-efficacy* as the writer's momentary belief in their ability to complete the given task independently, and *trust* as their momentary belief that the LLM will reliably support the task.

Our research questions are as follows:

- RQ1. What trajectory patterns emerge in users' self-efficacy and trust, and how do they interact over time?
- RQ2. How are users' prompting strategies associated with self-efficacy and trust trajectory patterns?
- RQ3. How are actual and perceived authorship associated with self-efficacy and trust trajectory patterns?

In this study, we focus on the trajectory patterns of user self-efficacy and trust rather than their static absolute scores for two primary reasons. First, self-efficacy and trust are highly subjective measures that vary across individuals as each person has a unique personal standard and interpretation of the scale [59, 60]. By examining within-person changes over time instead of a single reported score, we control for these individual differences. Second, LLM-assisted writing is a dynamic, iterative process where user perceptions evolve through interaction [44, 61, 62]. Trajectories capture this progressive aspect and allow us to examine how different prompting strategies and authorship experiences are associated with them. Below, we explain the details on how we conducted our study and analyzed the data. The study was reviewed and approved by the Institutional Review Board (IRB) of our institution prior to conducting the study.

3.1 Study Design

Here, we outline how we conducted our study.

3.1.1 Participants

We aimed to recruit a minimum of 300 participants after applying our exclusion criteria for quality control. This target was established following the criteria of VanVoorhis and Morgan [63], which recommend approximately 30 observations per cell to achieve 80% power in group comparisons. Our pilot study ($N = 20$) indicated the least frequent trajectory pattern occurred in approximately 10% of the sample, thus a sample of 300 would provide approximately 30 participants in the smallest group, meeting our power requirements.

We recruited a total of 313 initial participants through Prolific[1]. Since the task of our study was to write an argumentative essay in English, we used the preliminary filtering function of Prolific to only recruit native English speakers, ensuring consistent language proficiency across participants. We also

[1] https://www.prolific.com/

restricted recruitment to participants with approval ratings of 95% or higher, indicating high-quality participation in previous studies.

To observe participants' natural strategies regarding LLM use during the writing task, we recruited individuals with recent experience. A custom screening survey was used to confirm they had used an LLM within the last month. This qualification ensured that participants' use of the tool was based on preference or strategy, rather than a lack of familiarity. Participants who did not meet this criterion were screened out and received a nominal fee of 0.1 GBP.

Following data collection, we excluded 11 participants (3.5%) from analysis as they exceeded two hours to complete the task, or spent less than five minutes on writing as a quality control. One participant who took over two hours was excluded due to clear disengagement from the task. The remaining 10 participants spent less than five minutes on writing. We determined this duration insufficient to read the writing prompt, wait for AI responses to generate, review the output, and compose a response. This threshold fell below the 5th percentile (7.00 minutes) of writing durations, representing an inadequate timeframe for meaningful task completion. Beyond time-based criteria, our quality control also included reviewing participants' open-ended explanations for their ratings and post-survey responses (see Section 3.1.3) to identify off-topic or placeholder answers. However, no additional participants met these exclusion criteria. After applying all exclusion criteria, our final data consisted of 302 participants. We paid all eligible participants ($N = 313$) who completed our study 6.75 GBP ($\approx$ 9.11 USD), regardless of whether they were included in the final sample or not. The median study completion time was 39.21 minutes with the median writing time being 22.97 minutes.

To motivate participants to put their best effort into the writing task, we offered a performance-based incentive. Participants were informed that the top 10% of submissions would receive an additional 6.75 GBP based on essay quality, doubling their total payment. The submitted essays were graded using the College Board's scoring rubric and guidelines[2]. To support consistency in evaluation, we generated initial grades with the LLM, and one of the authors reviewed the LLM-generated scores alongside their brief justifications to adjust them when determining final grades. Overall, the author adjusted a total of 6.10% of the LLM-generated scores, and 34 participants (11.26%) received the additional bonus payment. We outline the detailed procedure as well as the prompt used in the Appendix 8.1.

3.1.2 Task

Participants were asked to complete an argumentative essay as the main task. The essay prompt was selected from the College Board AP English Language and Composition exam[3], similar to the setting of Siddiqui et al. [64]. We chose this exam for being a validated assessment with its standardized evaluation criteria, clear scoring guidelines, and accessibility for a general audience. Participants were asked to write a minimum of 300 words within a 30-minute time limit, adapted from the original exam setting. During the task, they could freely interact with the embedded ChatGPT from the interface (see Section 3.1.4) to generate ideas, draft content, or revise sentences.

3.1.3 Study Procedure

The study consisted of three main phases: pre-survey, writing task, and post-survey.

[2] https://apcentral.collegeboard.org/media/pdf/ap25-sg-english-language-set-1.pdf

[3] https://apcentral.collegeboard.org/media/pdf/ap25-frq-english-language-set-1.pdf

Firstly, participants completed a brief pre-survey measuring their initial, baseline self-efficacy and trust levels. They responded to two questions on 7-point Likert scales (1 = Not at all, 7 = Completely):

- [**Self-efficacy**] How confident are you in your ability to complete academic English writing tasks on your own?
- [**Trust**] How much do you trust LLMs to support you in completing academic English writing tasks?

These baseline ratings served dual purposes: establishing initial self-efficacy and trust scores for analysis, and providing participants with reference anchors for the in-the-moment ratings they would provide during the main task.

Secondly, participants proceeded to the main writing task as described in Section 3.1.2. To capture the dynamics of self-efficacy and trust, we collected in-the-moment ratings at the end of each interaction turn with the LLM. Before submitting a new prompt, participants were required to rate them on a 7-point Likert scale (1 = Not at all, 7 = Completely):

- [**Self-efficacy**] At this point, how confident are you in your ability to complete this writing task on your own?
- [**Trust**] At this point, how much do you trust the LLM to support you in completing this writing task?

Note that the same parallel questions about self-efficacy and trust from the pre-survey were repeatedly asked for each turn, allowing us to track their dynamic changes and treat both constructs as states that evolve throughout the interaction. The timing was chosen to allow participants sufficient opportunity to process the LLM's response and incorporate it into their draft, enabling more informed and reflective ratings.

The task concluded either when a participant met the 300-word requirement and chose to proceed, or automatically after 30 minutes. In either case, participants had to complete the ratings for their final interaction before moving on. Afterward, participants viewed a chronological timeline of their LLM interactions and provided brief open-ended explanations for each self-efficacy and trust rating they had given during the task.

Finally, participants completed a post-survey. They rated their perceived authorship of the final essay using two dimensions—ownership and agency—each measured on a 7-point Likert scale (1 = Not at all, 7 = Completely):

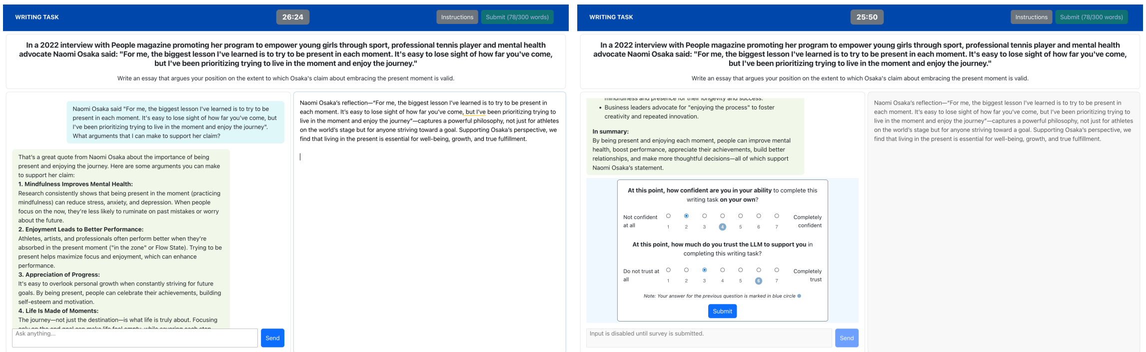
- [**Ownership**] How much do you feel like you are the author of the resulting text?
- [**Agency**] I felt like I was in control of the writing process during the task.

These questions were adapted from Qin et al. [42] and Draxler et al. [12], respectively. Lastly, participants were asked whether responding to in-the-moment ratings had influenced their interaction with the LLM, and, if so, in what ways.

We designed the rating questions to be as simple as possible to maintain natural interaction while still capturing the dynamics of self-efficacy and trust throughout the writing process. In the last post-survey question, 76.7% of participants responded that the rating process did not affect their interactions. However, 19.2% noted that it was either distracting or made them more mindful about their choices. We acknowledge this limitation in Section 6.

3.1.4 Study Interface

We deployed a custom web interface to conduct the study on Prolific, building on the design of Li et al. [6]. As shown in Figure 3.1a, the interface consisted of two primary panels: a chat panel on the left for interacting with the LLM, and a text editor on the right where participants could draft and revise their essay. The chat panel was implemented using vanilla `gpt-4.1`[4] to replicate a representative real-world usage environment. As explained in Section 3.1.3, users rated their in-the-moment self-efficacy and trust before sending the next prompt. For each interaction turn with LLM, the system mainly recorded the user's prompt, the LLM's response, the corresponding self-efficacy and trust ratings, the state of the essay at that moment, with timestamps captured at each logged action.



(a) Writing Interface (b) Self-Efficacy & Trust Ratings

Figure 3.1: Overview of the interface use in the study: (a) LLM-assisted writing workspace and (b) real-time self-efficacy and trust rating

Overview of the study platform showing two interfaces: (a) the writing workspace with the LLM chatbot and essay area, and (b) the rating interface for self-efficacy and trust.

To facilitate this rating process, the interface included two features. First, to help participants anchor their current feelings, a visual cue indicating their rating from the previous turn was displayed for each question. Second, to encourage deliberate reflection, both the prompt input and essay editor were disabled until the ratings were submitted.

Furthermore, to ensure data integrity and prevent gaming behaviors, the interface strictly prohibited URL manipulation to skip or revisit study phases and automatically saved participants' progress so that page refreshes would not affect task flow or recorded data. All refresh events were logged, and the authors manually reviewed these cases and confirmed that no participants displaying such noncompliant behaviors remained in the final data.

3.2 Data Analysis

We explain how we analyzed the data collected to answer each RQ. Our analyses consist of five main components: (1) identifying trajectory patterns of self-efficacy and trust (RQ1–3), (2) modeling their dynamic interactions using mixed-effects models (RQ1), (3) qualitatively coding user prompts to capture underlying intentions (RQ2), (4) analyzing prompting strategies through distribution of prompt intentions and their transition dynamics (RQ2), and (5) measuring actual authorship by comparing user

[4] https://platform.openai.com/docs/models/gpt-4.1

writing with LLM-generated content (RQ3). Other statistical comparisons between trajectory patterns are conducted as needed and reported in the results directly (Section 4).

3.2.1 Identifying Trajectory Patterns of Self-Efficacy and Trust

A central component of our analysis for all research questions is identifying trajectory patterns in self-efficacy and trust dynamics (RQ1–3). Here, we describe our analysis approach to classifying each participant's trajectories into distinct patterns.

To categorize these trajectories, we first established a statistically meaningful threshold for distinguishing reliable change from random measurement noise. Following prior works that similarly employed repeated administration of self-report scales to assess within-person change in psychological constructs [65, 66], we adopted the Minimal Detectable Change (MDC) threshold [67, 68] to distinguish reliable change from measurement error. At a 95% confidence interval, we found the theoretical threshold to be 2.080 points for self-efficacy and 2.037 points for trust. Given the discrete integer nature of the 7-point Likert scale, rounding the MDC up to 3 points would set a threshold that is disproportionately large relative to the scale's resolution, making it less sensitive to meaningful changes. Hence, we chose the nearest integer, 2 points, as the practical MDC cutoff indicating a significant change.

Consistent with established MDC-based practices to classify individuals into distinct groups [68, 69, 66], we used this threshold to first identify the initial trajectory patterns based on net change between final and initial scores:

- **Increase**: The final score was ≥ 2 points higher than the initial score.
- **Decrease**: The final score was ≥ 2 points lower than the initial score.
- **Stable**: All other cases where the final score was within ± 1 points of the initial score.

However, simple pre-post comparisons obscure meaningful within-session dynamics, as participants who appear stable in terms of net change may nevertheless exhibit qualitatively different intermediate trajectories during the interaction. To capture these dynamics, we further decomposed the *stable* pattern into three categories by examining whether any intermediate ratings crossed the same MDC-based 2-point threshold during the interaction:

- **Recovery**: The final score dropped ≥ 2 points but returned to within ± 1 point of initial.
- **Reversion**: The final score increased ≥ 2 points but returned to within ± 1 point of initial.
- **Stable**: All other cases which did not meet recovery or reversion criteria.

Through this process, we established five distinct trajectory patterns: *increase*, *decrease*, *stable*, *recovery*, and *reversion*. To maintain adequate power, we only included the patterns present in $\geq 10\%$ of participants in the main analyses.

3.2.2 Modeling the Dynamics of Self-Efficacy and Trust

To investigate how self-efficacy and trust interact over time (RQ1), we analyzed their potential bidirectional relationship using a linear mixed-effects model. This approach was chosen because our data contained repeated measurements from each participant across multiple interaction turns, violating the independence assumption of standard regression models.

Our primary analysis involved fitting two complementary models to examine the relationship from both directions: (1) one predicting *self-efficacy* scores from *turn*, *trust*, and their interaction, and (2) another one predicting *trust* scores from *turn*, *self-efficacy*, and their interaction. Here, *turn* was included as a key predictor in both models to capture how the relationship between self-efficacy and trust evolved over the course of the interaction.

We mean-centered all continuous predictors so that zero corresponded to the sample mean. Under this specification, each main effect represents the influence of a predictor when the interacting variable is fixed at its average level.

We tested two alternative random-effects structures. The first only included random intercepts, while the second included both random intercepts and random slopes for *turn*. Random intercepts account for individual baseline differences, while random slopes account for individual differences in the rate of change over time. A likelihood ratio test showed that the model with random slopes had a significantly better fit, so we report results from this specification.

3.2.3 Qualitative Analysis of User Prompts

Understanding users' prompting strategies (RQ2) requires first identifying the underlying intentions. We therefore qualitatively coded user prompts from the interaction logs. Each prompt was paired with the corresponding LLM response and ordered chronologically per participant to preserve the context.

The process was divided into three main steps. First, two authors independently reviewed 5.5% of the user prompts to identify recurring intentions. Here, LLM responses were visible during this step to ensure that prompts referring to prior outputs could be properly interpreted. Inter-rater reliability was high in this round ($\kappa = 0.867$). Any discrepancies or uncertainties were discussed, and a set of initial intention categories was established.

Second, the authors applied this scheme to an additional 10.1% of the data to validate its stability and agreement. No changes to the category list were required, and reliability remained substantially high ($\kappa = 0.789$).

Lastly, using the finalized categories, the remaining prompts were divided between the two authors and coded separately. Prompts could have multiple intentions when applicable. This process resulted in five main intention categories: *drafting*, *editing*, *ideating*, *information searching*, and *reviewing*. Table 3.1 summarizes these categories.

3.2.4 Analyzing Prompting Strategies

To understand user prompting strategies more holistically (RQ2), we performed two primary analyses based on the categorized prompt intentions. These analyses focused on investigating both the overall distribution of individual prompt intentions and the dynamic patterns of prompt transitions.

First, we analyzed the proportional usage of each prompting intention. Since participants completed different numbers of interaction turns, we normalized them by computing proportions relative to each person's total number of prompts. This measure allowed us to compare which intentions were strategically preferred across different trajectory patterns.

Second, to capture the dynamic and interactive nature, we analyzed transitions between consecutive prompting strategies. A static analysis of individual prompt proportions would not fully reveal how users' strategies evolved over the course of an interaction. We created a sequence of consecutive prompt pairs for each user (e.g., from prompt P_i to P_{i+1}). We then calculated the proportion of each transition type (e.g., from *drafting* to *editing*) relative to the total number of transitions. This analysis provided insight into the preferred strategic sequences of each user group, revealing patterns that individual prompting preferences cannot capture on their own.

Since there are a large number of possible transitions with five prompt categories, we only included those transitions exhibited by at least 10% of participants in our main analysis to maintain statistical power and robustness.

3.2.5 Measuring Actual Authorship

To analyze actual and perceived authorship (RQ3), we measured how participants incorporated LLM-generated content into their essays. While perceived authorship was captured through post-survey, actual authorship required an objective measurement of content overlap between LLM responses and user writings.

For measuring actual authorship, we employed two complementary approaches to capture different aspects of content incorporation: (1) lexical overlap and (2) semantic similarity. This combination was useful as direct copying (lexical overlap) and rephrasing (semantic similarity) are both common and distinct ways users integrate LLM-generated content into their work [70, 14, 40]. Lexical overlap was measured using ROUGE-3, which calculates recall-based tri-gram overlap by measuring how much of the reference text (LLM response) appears in the candidate text (user's writing). Semantic similarity was measured via text embeddings using `gemini-embedding-001`[5]. We captured conceptual overlap by computing the cosine similarity of high-dimensional vector embeddings from both the LLM response and user text.

We examined overall content adoption by comparing LLM-generated text across the entire interaction with the final essay. For each LLM response, we collected all preceding user messages to identify content the user had already contributed. We then extracted only the novel content which are originally generated by the LLM, excluding portions that restated user input since LLMs frequently rephrase user ideas. This LLM-original content and participants' final essays were tokenized and compared using both lexical and semantic similarity measures. This approach captures cases where users selectively adopted LLM content from various interaction points, regardless of when the adoption occurred.

[5]https://ai.google.dev/gemini-api/docs/embeddings

Table 3.1: Intention Categories and Definitions of User Prompts

Category	Subcategory	Description	Example
Drafting	Full Draft	User requests a complete essay based on provided criteria	Can you write a 300 word essay which argues for promoting this approach to empower young girls in sport.
	Partial Draft	User requests specific essay sections (thesis statements, paragraphs, sentences)	Give an introductory statement arguing for the point that many life coaches have advocated that embracing the present moment is a key belief to hold.
	Outline	User requests structural frameworks or organizational templates without content generation	Can you help me outline an argumentative essay on the validity of Naomi Osaka's claim?
Editing	—	User requests content revision, from targeted fixes to general improvements, including editing guidance	Can you add a bit about my travels to Japan and how mindfulness is extremely important in their culture and something surrounding ikigai?
Ideating	—	User seeks ideas, arguments, or assistance developing concepts and theses	Can you give me a few examples of how being present helps people in daily life?
Information Searching	Context	User requests background information about people, topics, or subjects for essay writing	Give me a brief background check on Naomi Osaka.
	Evidence	User requests specific statistics, quotes, or research to support particular arguments	Can you find a scientific study that shows the mental benefits of mindfulness practice?
	Writing	User requests writing strategies, techniques, or general guidance	What is the best way to structure an argumentative essay responding to this prompt?
Reviewing	—	User requests LLM assessment or evaluation of their writing without specific improvement directions	This is what I have so far. What do you think? [*Draft*]
Others	Language	User seeks assistance with spelling, grammar, word choice, synonyms, or phrasing	Synonym for savor.
	No Intent	User provides essay content in a follow-up prompt that continues from a previous request	[*Draft*]
	Others	User engages in conversation unrelated to essay writing	Thanks. I'm happy with this now

Chapter 4. Result

In this section, we first present descriptive statistics as an overview of the collected data and discuss each RQ in detail. We present our study results on users' self-efficacy and trust patterns and changes over time (RQ1), their association with prompting strategies (RQ2), and their relationship with actual and perceived authorship of the outcome (RQ3).

4.1 Descriptive Statistics

Our study collected a total of 1,410 user prompts and LLM response pairs. Each participant made an average of 4.67 turns ($SD = 3.87$, $min = 1.0$, $max = 25.0$), showing substantial variation and diversity in interaction patterns. Specifically, 368 prompts were labeled as *drafting* (26.1%), 305 as *editing* (21.6%), 185 as *ideating* (13.1%), 230 as *information searching* (16.3%), and 105 as *reviewing* (7.4%), with the remaining 217 prompts categorized as *others* (15.4%).

Participants' self-efficacy scores on a 7-point Likert scale had a mean of 5.07 ($SD = 1.74$, $min = 1$, $max = 7$), indicating moderate self-efficacy levels. The average change in self-efficacy per turn was -0.16 ($SD = 1.01$, $min = -6$, $max = 5$), suggesting that participants' self-efficacy decreased slightly with each interaction turn.

On the other hand, participants' trust levels in the LLM were higher than that of self-efficacy, with a mean of 6.00 ($SD = 1.28$, $min = 1$, $max = 7$). The average change in trust per turn was 0.14 ($SD = 0.72$, $min = -5$, $max = 4$), indicating that participants' trust in the LLM increased slightly with each interaction turn.

4.2 RQ1. What trajectory patterns emerge in users' self-efficacy and trust, and how do they interact over time?

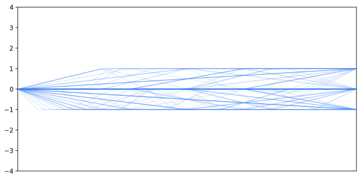
We first characterize the main trajectories of change in users' self-efficacy and trust during an interaction session with the LLM. Then, we examine how these two factors dynamically interact using linear mixed-effects regressions.

4.2.1 Trajectory Patterns of Self-Efficacy & Trust

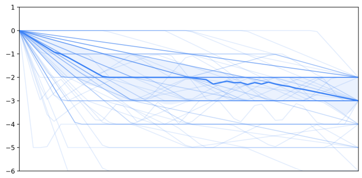
We categorized users' self-efficacy and trust into five trajectory patterns as explained in Section 3.2.1. Here, we first describe the distribution of participants across the five patterns. Then, we highlight the primary patterns for self-efficacy and trust to focus on later analyses, illustrated in Figure 4.1.

For self-efficacy, among 302 participants of our study, 56.6% ($N = 171$) were classified as *stable*, 27.5% ($N = 83$) as *decrease*, 10.9% ($N = 33$) as *recovery*, 4.0% ($N = 12$) as *increase*, and 1.0% ($N = 3$) as *reversion*. Overall, net decreases were more frequent than net increases. Hence, **we focus on three major trajectory patterns for our subsequent analyses of self-efficacy: *stable*** (Fig 4.1a), ***decrease*** (Fig 4.1b), **and *recovery*** (Fig 4.1c).

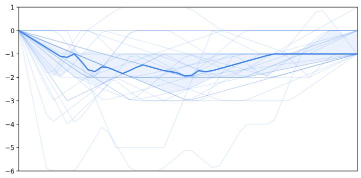
For each pattern, participants' distribution of self-efficacy scores varied; the *stable* group had a mean self-efficacy of 5.95 ($SD = 1.12$, $min = 2$, $max = 7$) with an average initial self-efficacy being



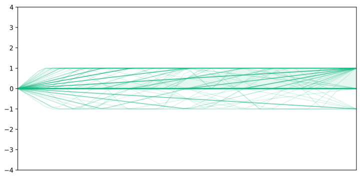
(a) Self-Efficacy — Stable ($N = 171$)



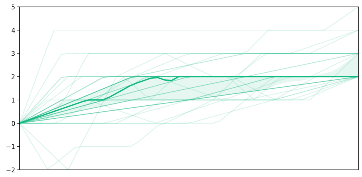
(b) Self-Efficacy — Decrease ($N = 83$)



(c) Self-Efficacy — Recovery ($N = 33$)



(d) Trust — Stable ($N = 231$)



(e) Trust — Increase ($N = 47$)

Figure 4.1: Normalized overlays of self-efficacy and trust trajectories for the five major trajectory patterns. The horizontal axis shows the time-normalized turn-by-turn progression from initial pre-survey rating (left) to final rating (right). The vertical axis shows the relative change in Likert-scale points after zero-centering all trajectories at each participant's pre-survey rating. Thin translucent lines represent individual participants, the solid line shows the group median, and the shaded band indicates the interquartile range.

6.04 ($SD = 1.07$, $min = 2$, $max = 7$), the *decrease* group had a mean self-efficacy of 3.70 ($SD = 1.67$, $min = 1$, $max = 7$) with an average initial self-efficacy being 6.00 ($SD = 1.00$, $min = 3$, $max = 7$), and the *recovery* group had a mean self-efficacy of 5.04 ($SD = 1.52$, $min = 1$, $max = 7$) with an average initial self-efficacy being 6.12 ($SD = 0.99$, $min = 3$, $max = 7$).

To determine whether these trajectories were driven by baseline differences, we compared these initial scores using a Kruskal–Wallis test. The test showed no significant differences among the groups ($p > 0.05$, not significant). This result shows that **participants' initial self-efficacy levels did not differ across the three main trajectories**, suggesting that subsequent decreases or recoveries were not attributable to baseline differences.

For trust, among the same 302 participants, 76.5% ($N = 231$) were classified as *stable*, 15.6% ($N = 47$) as *increase*, 5.3% ($N = 16$) as *recovery*, 1.7% ($N = 5$) as *reversion*, and 1.0% ($N = 3$) as *decrease*. Compared to self-efficacy, trust was predominantly stable, with net increases being more common than net decreases. Thus, **we focus on two major trajectory patterns for our subsequent analyses of trust: *stable*** (Fig 4.1d) **and *increase*** (Fig 4.1e).

Regarding score distributions, the *stable* group had a mean trust of 6.25 ($SD = 0.95$, $min = 3$, $max = 7$) with an average initial trust being 5.96 ($SD = 0.93$, $min = 3$, $max = 7$), whereas the *increase* group had a mean trust of 5.62 ($SD = 1.45$, $min = 1$, $max = 7$) with a notably lower average initial trust being 4.00 ($SD = 1.16$, $min = 1$, $max = 5$).

To validate this disparity, we compared the initial trust scores between the two groups using a Mann–Whitney U test. The test revealed a significant difference ($U = 9876$, $p < 0.001$), showing that **trust increases occurred primarily among participants who began with lower initial trust**, whereas those who with higher trust remained stable.

4.2.2 Interaction Between Self-Efficacy and Trust Over Time

We present the results of mixed-effects models examining how self-efficacy and trust influence each other. The details of the regression models are explained in Section 3.2.2. We first report changes in self-efficacy over turns, followed by trust dynamics.

Changes in Self-Efficacy Through the mixed-effects model, we first found a significant main effect for *Turn* on self-efficacy ($\beta = -0.115$, $SE = 0.023$, $z = -4.90$, $p < 0.001$). On average, **participants' self-efficacy decreased over turns**, specifically, 0.115 points per turn on the 7-point Likert scale. We also found a significant main effect for *Trust* on self-efficacy ($\beta = 0.100$, $SE = 0.038$, $z = 2.63$, $p < 0.01$). This result suggests that **participants with a higher level of trust also maintained a higher overall self-efficacy** throughout the interactions.

Furthermore, we found a significant *Turn* $\times$ *Trust* interaction ($\beta = 0.038$, $SE = 0.009$, $z = 4.10$, $p < 0.001$). This indicates that the effect of turns on self-efficacy was moderated by trust. In other words, **for participants with higher trust, the decrease in self-efficacy over turns was significantly smaller**, showing more resilient trajectories. This suggests that a high trust level may act as a buffer, mitigating the decrease in self-efficacy as the interaction progresses.

Changes in Trust In a complementary mixed-effects model, we found a significant main effect for *Turn* ($\beta = 0.111$, $SE = 0.011$, $z = 10.13$, $p < 0.001$) on trust. Unlike self-efficacy, the result suggests that **participants' trust increased over turns**, specifically, 0.111 points per turn on the 7-point Likert scale.

However, the effect of *self-efficacy* on trust was not significant ($p > 0.05$, not significant). Nevertheless, in the *Turn* $\times$ *self-efficacy* interaction, we found a significance ($\beta = 0.017$, $SE = 0.004$, $z = 3.67$, $p < 0.001$). This result shows that the effect of turns on trust was moderated by self-efficacy. Although self-efficacy was not a direct predictor of trust, **high self-efficacy showed a faster increase in trust over turns**, while lower self-efficacy showed a more gradual increase.

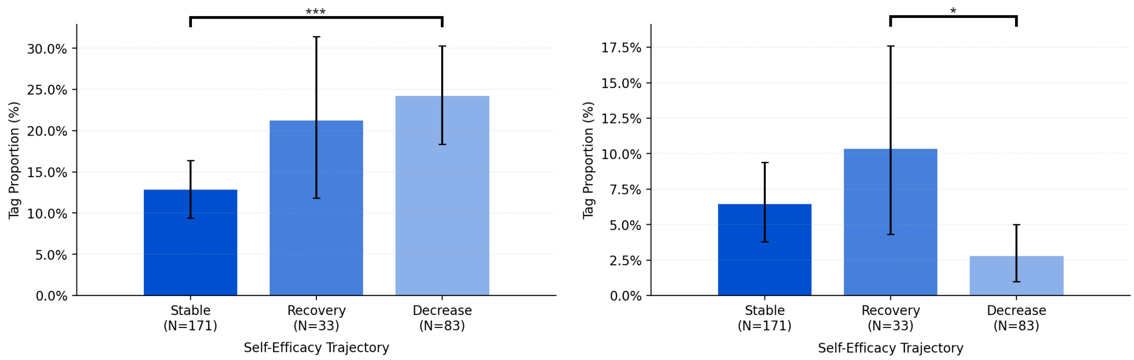
4.3 RQ2. How are users' prompting strategies associated with self-efficacy and trust trajectory patterns?

We analyzed how self-efficacy and trust trajectory patterns relate to users' prompting strategies across two dimensions: (1) individual prompt usage and (2) prompt-to-prompt transitions, as explained in Section 3.2.3 and 3.2.4 respectively. To further contextualize these trajectory-level findings, we also conducted supplementary turn-level analyses of self-efficacy and trust across prompt categories. We present these additional findings in Appendix 8.2.

4.3.1 Individual Prompt Usage

Self-Efficacy We conducted Kruskal-Wallis tests to compare the proportional use of five main categories of prompts (*drafting*, *ideating*, *editing*, *information searching*, *reviewing*) across three self-efficacy patterns (*stable*, *decrease*, *recovery*). The result showed significant differences for *editing* ($H(2) = 13.57$, $p < 0.001$) and *reviewing* ($H(2) = 6.71$, $p < 0.05$). No significant differences were observed for the remaining three categories ($p > 0.05$, not significant).

We further conducted post-hoc comparisons for significant prompts using Dunn's test with Holm-Bonferroni correction [71] for multiple comparisons. We illustrate the pairwise comparison results of self-efficacy in Figure 4.2.



(a) *Editing* Prompt (b) *Reviewing* Prompt

Figure 4.2: Proportions of *editing* and *reviewing* prompts across different self-efficacy trajectory patterns. ($*p < 0.05$, $***p < 0.001$)

We found that users with *decrease* self-efficacy trajectory had a significantly higher proportion of *editing* prompts than those in *stable* trajectory ($p_{adj} < 0.001$, Fig 4.2a). Moreover, users with *recovery* trajectory showed a significantly higher proportion of *reviewing* prompts than those in *decrease* trajectory ($p_{adj} < 0.05$, Fig 4.2b). The other comparisons were not significant after correction. In other words, our results show that **users in *decrease* trajectory request more for editing than those in *stable*, while using fewer requests for reviewing than those in *recovery*.**

Trust We conducted Mann-Whitney U tests to compare the prompt usage across two trust trajectory patterns (*stable*, *increase*). We found that **users in *increase* trajectory request significantly more information searching than those in *stable* trajectory** ($U = 4293.5$, $p < 0.05$, Fig 4.3). No significant differences were found for others.

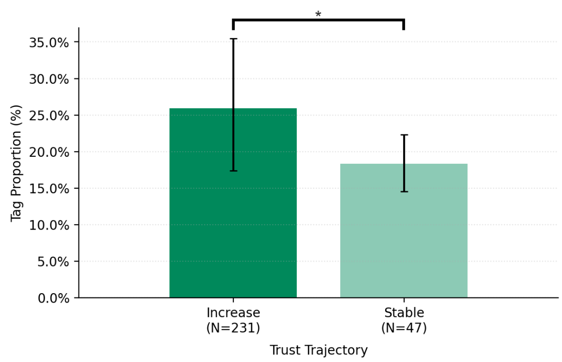
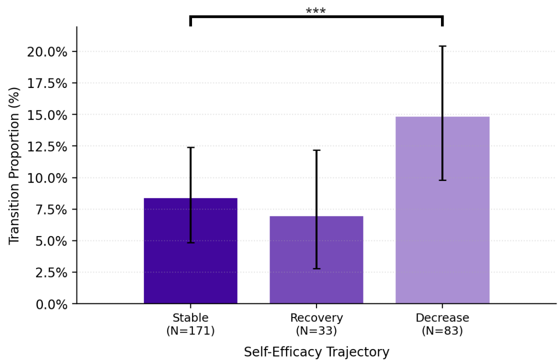


Figure 4.3: Proportions of *information searching* prompts across different trust trajectory patterns. ($*p < 0.05$)

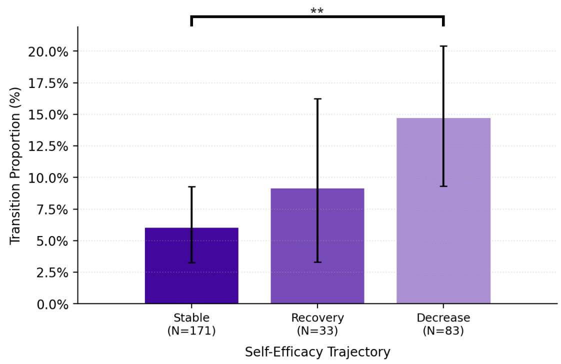
4.3.2 Prompt-to-Prompt Transitions

Self-Efficacy We conducted Kruskal-Wallis tests across self-efficacy patterns and found a significant difference for the *drafting—editing* transition ($H(2) = 16.55$, $p < 0.001$) and *editing—editing* transition ($H(2) = 9.24$, $p < 0.01$).

Post-hoc Dunn's test with Holm-Bonferroni correction [71] showed that **users with *decrease* self-efficacy trajectory showed a significantly higher proportion of *drafting—editing* transitions than those in the *stable* trajectory** ($p_{adj} < 0.001$, Fig 4.4a). Similarly, **they also had a significantly higher proportion of *editing—editing* transitions than the *stable* trajectory** ($p_{adj} < 0.01$, Fig 4.4b). Other pairwise comparisons were not significant. We illustrate these results in Figure 4.4. These findings reveal that users with decreasing self-efficacy exhibit a characteristic strategy of immediately transitioning from generation to editing, followed by consecutive editing behaviors.



(a) *Drafting—Editing* Transition



(b) *Editing—Editing* Transition

Figure 4.4: Proportion of *generating—editing* and *editing—editing* transition across different self-efficacy trajectory patterns. (**$p < 0.01$, ***$p < 0.001$)

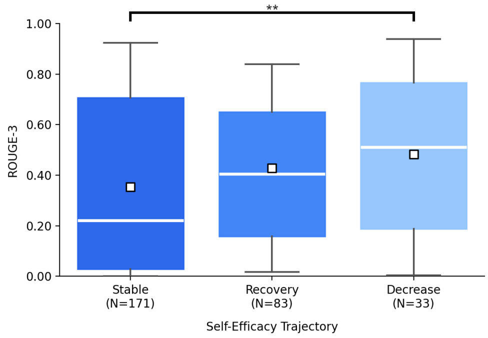
Trust We conducted Mann-Whitney U tests for trust patterns across filtered transition types. However, we did not find any significant differences for any significant transition patterns ($p > 0.05$, not significant).

4.4 RQ3. How are actual and perceived authorship associated with self-efficacy and trust trajectory patterns?

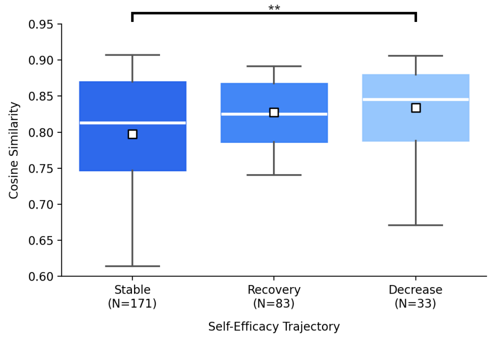
We analyzed the relationship between the trajectory patterns and the authorship of the outcome. Note that actual authorship of each user was measured via lexical overlap and semantic similarity, and perceived authorship was measured via post-survey on ownership and agency, as explained in Section 3.2.5.

Self-Efficacy We conducted Kruskal-Wallis tests across three self-efficacy patterns (*stable*, *decrease*, *recovery*) and found significant differences between them in terms of lexical overlap ($H(2) = 12.65$, $p < 0.01$), semantic similarity ($H(2) = 12.77$, $p < 0.01$), ownership ($H(2) = 59.28$, $p < 0.001$), and agency ($H(2) = 25.93$, $p < 0.001$).

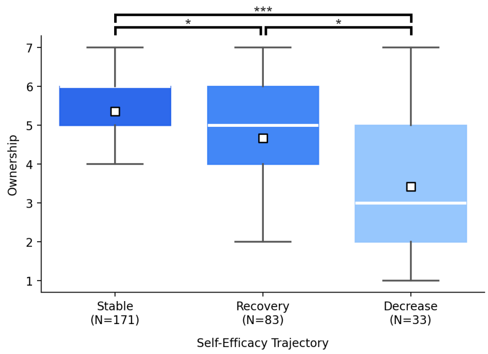
We further conducted post-hoc comparisons using Dunn's test with Holm-Bonferroni correction [71] to correct for multiple comparisons. We illustrate the consequent pairwise comparison results in Figure



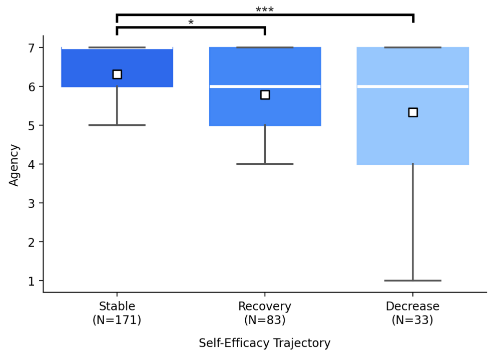
(a) Lexical Overlap



(b) Semantic Similarity



(c) Ownership



(d) Agency

Figure 4.5: Actual (a–b) and perceived (c–d) authorship across different self-efficacy trajectory patterns. (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$)

4.5.

For actual authorship, we found the *decrease* trajectory had significantly higher lexical overlap than *stable* ($p_{adj} < 0.01$, Fig 4.5a). The other pairs were not significant after correction. Similarly, *decrease* trajectory had significantly higher semantic similarity than *stable* ($p_{adj} < 0.01$, Fig 4.5b). The other pairs were not significant.

For perceived authorship, we found the ownership significantly differed across all pairs (Fig 4.5c). Specifically, there was a clear hierarchy of *stable* > *recovery* ($p_{adj} < 0.05$), *recovery* > *decrease* ($p_{adj} < 0.05$), and *stable* > *decrease* ($p_{adj} < 0.001$). We found a similar pattern in agency (Fig 4.5d); *stable* trajectory had significantly higher agency than both *recovery* ($p_{adj} < 0.05$) and *decrease* ($p_{adj} < 0.001$). The *recovery*—*decrease* pair was not significant.

Across the analyses, we observed a **consistent significant gap between self-efficacy trajectories of *decrease* and *stable*, where users in *decrease* trajectory show lower values for both actual and perceived authorship**. The *recovery* trajectory was intermediate, however, the differences were relatively smaller and often non-significant.

Trust We conducted Mann-Whitney U tests across two trust patterns (*stable*, *increase*) and found no noticeable differences between them ($p > 0.05$, not significant) across every measure of authorship.

Chapter 5. Discussion

In this section, we first interpret our findings. We then propose design implications for understanding and supporting authorship in human-LLM collaboration. Lastly, we discuss the generalizability of our study results.

5.1 Interpretation of Results

Here we discuss our interpretation of the results and their implications.

5.1.1 Diverging Dynamics of Self-Efficacy and Trust

We found that participants' self-efficacy decreased while their trust increased over the course of the interaction (Section 4.2.2). The decline in self-efficacy can be attributed to upward social comparison with the model's consistently high-quality outputs [18]. When users repeatedly observed the system producing polished text more quickly than they could themselves, the contrast may have highlighted gaps in their own abilities. Prior work shows that performing alongside superior agents can erode self-efficacy through such comparisons, especially when individuals have limited opportunities to demonstrate their own competence [72]. LLMs may amplify this dynamic by generating complete revisions instantly, leaving little room for users to formulate or refine text themselves. Even when users provide conceptual direction, the model executes the improvement, potentially reducing opportunities to develop skills or attribute outcomes to their own efforts.

On the other hand, the increase in trust aligns with outcome-based trust calibration in automation [27, 73], where trust evolves through the repeated observation of successful performance. In such calibration processes, users adjust trust by identifying reasoning flaws or system inaccuracies [74, 75]. In writing, however, there is no binary correct answer, and evaluating revision quality requires significant cognitive effort. Hence, users rely on surface-level cues such as fluency and coherence [76, 77], with each fluent output perceived as a success. This leads users to overweight the model's apparent competence, driving trust accumulation. These dynamics create an asymmetric structure: users receive frequent confirmation of the system's competence but little confirmation of their own. This imbalance may explain why trust and self-efficacy diverged during collaboration, suggesting that the capabilities making LLMs powerful writing partners may simultaneously shift users' perceptions of their own competence and contribution.

5.1.2 Role of Prompting Strategies in Self-Efficacy Trajectories

Our findings show that users whose self-efficacy declined relied heavily on *editing* requests while those who recovered self-efficacy used more *review* requests (Section 4.3). This pattern can be understood through findings from the educational domain, showing that direct corrective feedback positions learners as passive recipients, while diagnostic feedback that identifies issues maintains learner agency and supports self-efficacy [78]. These studies emphasize that effective feedback should provide users the ability to maintain control over revision decisions in determining when, where, and how to implement suggested changes.

Heavy reliance on editing requests reflects a dynamic where users position the AI as the authoritative agent responsible for text improvement. This can reduce their own agency as the opportunities to make revision decisions themselves become limited. On the other hand, the shift toward review-oriented interactions indicates users who maintain their position as primary authors while leveraging the AI as a diagnostic tool. This approach preserves writer agency, consistent with research showing that peer review reduces anxiety while building self-efficacy [79]. Therefore, understanding how users perceive the AI's role as a collaborator is important, and our work shows how these perceptions can be understood through their prompting behaviors.

5.2 Design Implications

We present design implications for understanding and supporting authorship in human-LLM collaboration.

5.2.1 Conceptualizing Self-Efficacy as a Dynamic Process

Our findings suggest that self-efficacy operates as a dynamic process during human-LLM collaboration rather than a static trait. Prior research on self-efficacy and authorship has shown that higher self-efficacy aligns with stronger authorship and authorial identity [80, 81], whereas lower self-efficacy is associated with greater susceptibility to plagiarism [82]. However, existing work largely treats self-efficacy as a static, between-person attribute that is typically measured at a single point in time.

In contrast, our findings reveal that within-session trajectories of self-efficacy provide a more informative basis for understanding authorship loss. While participants' initial self-efficacy scores did not differ across trajectory groups (Section 4.2.1), within-session declines and fluctuations in self-efficacy were significantly associated with decreased authorship (Section 4.4). Specifically, users who experienced declining self-efficacy not only lost actual authorship by adopting more of the LLM's direct phrasing and ideas but also reported weaker perceived ownership and agency over their final text. Moreover, users whose self-efficacy recovered showed intermediate authorship outcomes between the stable and decreasing groups, a pattern that a simple pre–post comparison would obscure.

Taken together, these findings suggest that understanding authorship in human–LLM collaboration requires accounting for how self-efficacy evolves over the course of writing. Conceptualizing self-efficacy as a dynamic trajectory offers a more nuanced lens for examining when and why authorship is strengthened or weakened, and provides a basis for assessments that can, in turn, inform the design of interventions to preserve user authorship.

5.2.2 Using Behavioral Signals to Examine Self-Efficacy Shifts

Self-efficacy decline can undermine both actual and perceived authorship (Section 4.4), making it important to understand when and how these shifts occur. Our findings suggest that user behavior during collaboration can serve as unobtrusive signals of changes in self-efficacy, helping reveal moments in which authorship may become vulnerable. Specifically, we identified two key behavioral patterns: prompting strategies and how users incorporate LLM outputs.

Prompting strategies strongly correspond to self-efficacy trajectories (Section 4.3). A high frequency of consecutive editing prompts or rapid shifts from drafting to editing may indicate a decline in self-efficacy, whereas an increase in reviewing prompts may reflect efforts to regain a sense of control. Such shifts can be detected using methods like prompt intent classification [83], providing a way to identify moments of potential authorship loss.

Moreover, how users integrate LLM outputs provides complementary insight into self-efficacy shifts. As self-efficacy declines, users tend to adopt a larger portion of the model's suggestions (Section 4.4). Tracking how semantic similarity or lexical overlap between user-written text and model-generated text changes over time can reveal shifts from active authorship toward passive acceptance. Techniques such as keystroke logging [84] enable capturing these incorporation dynamics, offering additional evidence of when users' contributions begin to recede.

Together, these signals offer a foundation for examining self-efficacy trajectories in situ and open up opportunities to explore strategies that help sustain authorship when self-efficacy is at risk of declining.

5.2.3 Opportunities for Supporting Authorship in LLM-Assisted Writing

Our findings on behavioral signals and self-efficacy trajectories open up several directions for designing systems that better support user authorship in LLM-assisted writing. We discuss opportunities for designing environments that can foster a more intentional and reflective collaboration, helping users maintain active agency rather than drifting into passive acceptance.

One opportunity lies in helping users stay aware of their interaction patterns and shift toward more reflective engagements. Our results revealed users with declining self-efficacy fell into repetitive "editing loops" of requesting direct revisions (Section 4.3.2), while those regaining self-efficacy sought evaluation rather than relying on the model to fix their text (Section 4.3.1). Since prior work suggests that making people's own thoughts visible can foster metacognitive awareness [85], systems could help users recognize their engagement patterns by making collaboration history more visible [86] or steering responses toward review and reflection [87]. Such interventions may help users move from passive acceptance of suggestions toward more intentional collaboration, with greater agency over their writing.

Another opportunity lies in supporting users' information needs during the writing process. Our results show that information-seeking behaviors were associated with increasing trust in the LLM (Section 4.3.1), which in turn buffered against declines in self-efficacy (Section 4.2.2). Notably, trust increases occurred primarily among participants who began with lower initial trust (Section 4.2.1), suggesting that informational support may be especially beneficial during early stages. Building on these findings, systems could provide relevant background information or supporting evidence when users need it, an approach shown to improve writing quality [88] while keeping users in control. However, as LLMs may occasionally retrieve irrelevant content or hallucinate sources [89, 90], systems could consider strategies to maintain calibrated trust, such as signaling uncertainty [91] or prompting users to validate sources [92].

5.3 Generalization of Results

Our study examined how users' self-efficacy and trust evolve during LLM-assisted writing. We found that while trust in the LLM increased over time, users' self-efficacy decreased. We discuss how this pattern can generalize to other cognitively demanding, open-ended tasks where LLMs can be particularly valuable in providing diverse assistance.

Prior work on AI-assisted tools, particularly in creative domains, has largely focused on how LLMs can enhance efficiency [93], performance [94], and user satisfaction [95]. However, our findings point to a different dimension of AI-assisted work, indicating that users' self-efficacy in their abilities and sense of authorship may decline during collaboration with LLMs. This highlights the need for future research to examine not only what AI tools can produce, but also how they affect users' psychological states. These considerations may become especially salient in multimodal AI systems, where distinguishing individual contributions is more challenging than in text-based collaboration.

Our findings are also particularly significant in educational contexts, where the consequences of declining self-efficacy and losing authorship can be extremely severe. In learning environments, students' belief in their own abilities is fundamental to academic motivation and long-term development. Recent work has already shown that AI overreliance can diminish students' critical thinking and engagement [53]. Our observed decline in self-efficacy may represent a crucial pathway to these detrimental outcomes, limiting students' capacity to approach challenging tasks independently.

Both creative and educational contexts are examples of a broader category of cognitively demanding tasks where this tension between increased performance and diminished self-efficacy is likely to emerge. Thus, our results and implications can be extended beyond these domains to collaborative AI that supports cognitively intensive work.

Chapter 6. Limitations and Future Work

We acknowledge several limitations of our study and suggest potential future work.

First, our measures of self-efficacy and trust relied on self-reported ratings collected after each interaction turn with the LLM. Although most participants indicated that this process did not affect their interactions, a subset of the participants reported that it was distracting and made them more self-conscious about their choices (Section 3.1.3). Despite our efforts to design the ratings to be as lightweight as possible, this suggests that it may still have influenced participants' natural interaction patterns in subtle ways, even among those who reported no disruption. Specifically, participants who experienced the ratings as distracting may have partially shifted their work to external tools (e.g., other LLMs), but because our logging was limited to in-platform events (Section 3.1.4), we cannot determine whether this occurred. To better capture users' natural interactions, future work could explore less intrusive approaches, such as inferring self-efficacy and trust from behavioral signals—including typing latency and editing pauses [96], dwell time [97], or eye gaze [98]. Alternatively, expanding data collection to include pageview and cross-platform activity could enable post-hoc filtering of sessions involving external tool use, improving the integrity of the data.

Second, our findings identified empirical patterns in how self-efficacy and trust evolved within an interaction session with the LLM, but not the qualitative reasons behind them. To uncover why these shifts occur, future work could employ qualitative methods like user interviews or think-aloud protocols to capture the moment-to-moment rationale behind users' decisions. Bridging the gap between observed behaviors and subjective goals would provide actionable guidance for designing effective interventions to preserve user authorship (Section 5.2.3). Furthermore, although we analyzed patterns within a multi-turn interaction, the study was confined to a single session. This limits our understanding of whether the observed shifts are temporary or develop into long-term changes. To capture how these dynamics change, future research could extend our design to a longitudinal study that tracks user perceptions across multiple sessions, revealing how repeated interactions shape users' evolving relationship with AI assistance.

Finally, this study did not account for individual or system-related factors that prior work has shown to influence self-efficacy and trust, such as task proficiency [99] or variations in system tone and expression [100, 101]. Our analysis also used a single LLM, so we could not examine how such stylistic and behavioral differences across models might alter user perceptions [102]. Future research could extend this work by studying more diverse user groups and varying LLM configurations to better understand how these contextual factors interact with the dynamics of self-efficacy and trust.

Chapter 7. Conclusion

We investigated how self-efficacy and trust evolve during LLM-assisted writing through an empirical study with 302 participants. We found that while trust generally increases throughout the interaction, self-efficacy tends to decline, buffered by high trust. Trajectory-level analyses revealed that users with declining self-efficacy were more likely to rely on passive editing strategies and showed diminished authorship, whereas those with recovering self-efficacy engaged more in review-oriented interactions with intermediate authorship outcomes between the declining and stable groups. Our findings show that self-efficacy operates as a dynamic process that critically shapes how people collaborate with LLMs, highlighting the need to examine and respond to within-session shifts to preserve user agency and authorship. We call for future work to move beyond single-session self-reports, exploring longitudinal studies and unobtrusive measures to better understand the long-term trajectories of self-efficacy, trust, and human agency in AI-assisted work.

Chapter 8. Appendix

8.1 Procedure for Essay Grading

We generated initial essay scores using a prompt adapted from the automated essay scoring prompts of Yoo et al. [103] with the official College Board rubric for the AP English Language and Composition Argument Essay[1]. The rubric evaluates three criteria for a total of six points: (1) thesis (0-1), (2) evidence and commentary (0-4), and (3) sophistication (0-1). We instructed `gpt-5-nano`[2] to generate scores with brief justifications for each criterion, enabling the author to review and finalize the grades. The detailed prompt is presented below.

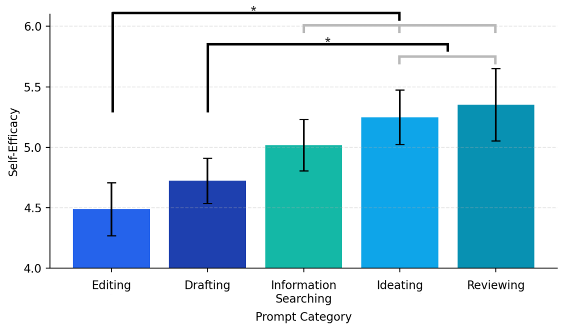
```
You are grading an AP English Language and Composition Argument Essay.

### Rubric:
<rubric_criteria>

Comments: For each row, provide a justification of the score in this exact bullet format:
- Thesis: [your justification here]
- Evidence & Commentary: [your justification here]
- Sophistication: [your justification here]

### Output Instructions:
Return ONLY a valid JSON object with the following keys:
{thesis: integer (0-1), evidence_commentary: integer (0-4), sophistication: integer (0-1), comment: string}

### Essay Prompt:
<prompt_text>

### Student Essay:
<essay_text>
```

After obtaining the initial rubric-based scores, one of the authors iteratively reviewed the essays and adjusted these scores to determine final bonus recipients. Since many essays received similar high scores, each iteration involved manually evaluating top-ranked essays, adjusting scores to better differentiate quality, and re-ranking all entries. This process was repeated until the top-performing subset was clearly distinguished and stabilized. In total, the author adjusted 6.10% of the LLM-generated scores, and the top 34 participants (11.26%) received the bonus payment.

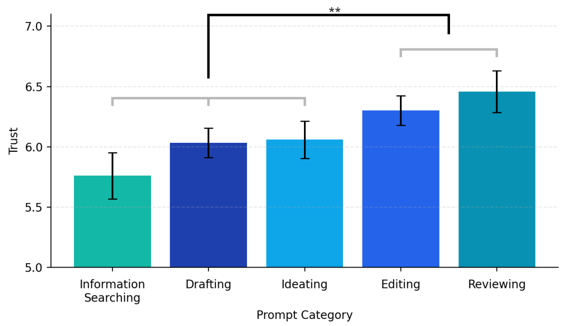
[1] https://apcentral.collegeboard.org/media/pdf/ap25-sg-english-language-set-1.pdf
[2] https://platform.openai.com/docs/models/gpt-5-nano

8.2 Turn-Level Analysis of Self-Efficacy and Trust by Prompt Category

This appendix reports supplementary turn-level analyses examining how users' self-efficacy and trust ratings differed across prompt categories during the interaction. These analyses provide additional context for interpreting the trajectory-level findings of Section 4.3 by revealing how each prompt type relates to users' momentary ratings at specific turns.



(a) Self-Efficacy by Prompt Category



(b) Trust by Prompt Category

Figure 8.1: Turn-level self-efficacy and trust ratings across prompt categories (*$p < 0.05$, **$p < 0.01$)

8.2.1 Self-Efficacy

We conducted a Kruskal–Wallis test to compare turn-level self-efficacy across the five main prompt categories and found significant differences among them ($H(4) = 28.74$, $p < 0.001$). Post-hoc Dunn tests with Holm-Bonferroni correction [71] revealed that *editing* prompts were associated with significantly lower self-efficacy than *ideating*, *information searching*, and *reviewing* prompts ($p_{adj} < 0.05$ for all). Moreover, *drafting* prompts also showed lower self-efficacy than *ideating* and *reviewing* ($p_{adj} < 0.05$ for all). We illustrate this result in Figure 8.1a.

These turn-level patterns align with our trajectory-level findings in Sections 4.3.1 and 4.3.2. Users in the *decrease* trajectory relied heavily on *editing* prompts and engaged in *drafting-to-editing* transitions, while those in the *recovery* trajectory made more *reviewing* requests. The turn-level results suggest that *editing* and *drafting* are more often used when users feel less efficacious, whereas *ideating*, *information searching*, and *reviewing* tend to occur when self-efficacy is relatively higher. This indicates that heavy reliance on editing both reflects and potentially reinforces lower-efficacious trajectories, while reviewing appears characteristic of users who maintain or regain their sense of capability.

8.2.2 Trust

We also conducted a Kruskal–Wallis test for trust ratings across prompt categories and found significant differences ($H(4) = 39.02$, $p < 0.001$). Post-hoc Dunn tests with Holm-Bonferroni correction [71] revealed that *reviewing* and *editing* prompts were associated with higher trust than the other three ($p_{adj} < 0.01$ for all), while *information searching* showed the lowest mean trust (M=*5.76*). We illustrate this result in Figure 8.1b.

These offer contextual insight into the trajectory-level findings of Section 4.2.1 and 4.3.1. Users in the *increase* trajectory began with lower initial trust and more frequently used *information searching* prompts, and the turn-level analysis similarly shows that *information searching* turns coincide with lower trust. In contrast, *editing* and *reviewing* prompts, which involve working more directly with the outcome text, are associated with higher trust. Taken together, these patterns suggest that users tend to deal with lower trust through information-seeking interactions, and that once trust is higher, they are more likely to request edits and reviews that directly shape the outcome text.

8.2.3 Considerations for Interpretation

It is important to acknowledge that this turn-level analysis aggregates data across all interactions without normalizing for the varying number of turns per participant. Consequently, participants who engaged in more turns contribute proportionally more data points to each prompt category. As a result, these findings may be influenced by the individual traits of highly active users. While this analysis offers supportive evidence that complements our trajectory-level findings, we emphasize that these associations serve as an exploratory context for interpreting the trajectory patterns rather than definitive evidence regarding the intrinsic properties of each prompt category.

Bibliography

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.

[3] M. Lee, P. Liang, and Q. 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*, pp. 1–19, 2022.

[4] N. Sommers, "Revision strategies of student writers and experienced adult writers," *College Composition & Communication*, vol. 31, no. 4, pp. 378–388, 1980.

[5] L. Flower and J. R. Hayes, "A cognitive process theory of writing," *College Composition & Communication*, vol. 32, no. 4, pp. 365–387, 1981.

[6] Z. Li, C. Liang, J. Peng, and M. Yin, "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*, pp. 1–25, 2024.

[7] P. S. Dhillon, S. Molaei, J. Li, M. Golub, S. Zheng, and L. P. Robert, "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*, pp. 1–18, 2024.

[8] M. Reza, N. M. Laundry, I. Musabirov, P. Dushniku, Z. Y. u, K. Mittal, T. Grossman, M. Liut, A. Kuzminykh, and J. J. Williams, "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*, pp. 1–18, 2024.

[9] S. Mysore, D. Das, H. Cao, and B. Sarrafzadeh, "Prototypical human-ai collaboration behaviors from llm-assisted writing in the wild," *arXiv preprint arXiv:2505.16023*, 2025.

[10] A. Guo, S. Sathyanarayanan, L. Wang, J. Heer, and A. X. Zhang, "From pen to prompt: how creative writers integrate ai into their writing practice," in *Proceedings of the 2025 Conference on Creativity and Cognition*, pp. 527–545, 2025.

[11] A. Nguyen, Y. Hong, B. Dang, and X. Huang, "Human-ai collaboration patterns in ai-assisted academic writing," *Studies in Higher Education*, vol. 49, no. 5, pp. 847–864, 2024.

[12] F. Draxler, A. Werner, F. Lehmann, M. Hoppe, A. Schmidt, D. Buschek, and R. Welsch, "The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors," *ACM Transactions on Computer-Human Interaction*, vol. 31, no. 2, pp. 1–40, 2024.

[13] A. H.-C. Hwang, Q. V. Liao, S. L. Blodgett, A. Olteanu, and A. Trischler, "'it was 80% me, 20% ai': Seeking authenticity in co-writing with large language models," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–41, 2025.

[14] C. K. Y. Chan, "Is ai changing the rules of academic misconduct? an in-depth look at students' perceptions of'ai-giarism'," *arXiv preprint arXiv:2306.03358*, 2023.

[15] P. Spitzer, J. Holstein, P. Hemmer, M. Vössing, N. Kühl, D. Martin, and G. Satzger, "Human delegation behavior in human-ai collaboration: The effect of contextual information," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–28, 2025.

[16] M. Schemmer, N. Kuehl, C. Benz, A. Bartos, and G. Satzger, "Appropriate reliance on ai advice: Conceptualization and the effect of explanations," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422, 2023.

[17] X. J. Yang, C. Schemanske, and C. Searle, "Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation," *Human Factors*, vol. 65, no. 5, pp. 862–878, 2023.

[18] A. Bandura, "Self-efficacy: toward a unifying theory of behavioral change.," *Psychological review*, vol. 84, no. 2, p. 191, 1977.

[19] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.

[20] A. Bandura, *Self-efficacy: The exercise of control*, vol. 11. Freeman, 1997.

[21] D. H. Schunk, "Self-efficacy and academic motivation," *Educational psychologist*, vol. 26, no. 3-4, pp. 207–231, 1991.

[22] A. D. Stajkovic and F. Luthans, "Self-efficacy and work-related performance: A meta-analysis.," *Psychological bulletin*, vol. 124, no. 2, p. 240, 1998.

[23] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[24] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma, "Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023.

[25] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (New York, NY, USA), p. 295–305, Association for Computing Machinery, 2020.

[26] B. Lubars and C. Tan, "Ask not what ai can do, but what ai should do: Towards a framework of task delegability," *Advances in neural information processing systems*, vol. 32, 2019.

[27] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[28] H. Kumar, I. Musabirov, M. Reza, J. Shi, X. Wang, J. J. Williams, A. Kuzminykh, and M. Liut, "Guiding students in using llms in supported learning environments: Effects on interaction dynamics, learner performance, confidence, and trust," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–30, 2024.

[29] G. He, L. Kuiper, and U. Gadiraju, "Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–18, 2023.

[30] J. Li, Y. Yang, Q. V. Liao, J. Zhang, and Y.-C. Lee, "As confidence aligns: Understanding the effect of ai confidence on human self-confidence in human-ai decision making," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2025.

[31] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proceedings of the ACM on Human-computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[32] S. Ma, X. Wang, Y. Lei, C. Shi, M. Yin, and X. Ma, ""are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024.

[33] S. C. Kohn, E. J. De Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Frontiers in psychology*, vol. 12, p. 604977, 2021.

[34] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS quarterly*, pp. 189–211, 1995.

[35] X. Wang, S. Tulk Jesso, S. Kojaku, D. M. Neyens, and M. S. Kim, "Viztrust: A visual analytics tool for capturing user trust dynamics in human-ai communication," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2025.

[36] W. Du, V. Raheja, D. Kumar, Z. M. Kim, M. Lopez, and D. Kang, "Understanding iterative revision from human-written text," *arXiv preprint arXiv:2203.03802*, 2022.

[37] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

[38] Z. Guo, Y. Wu, J. D. Hartline, and J. Hullman, "A decision theoretic framework for measuring ai reliance," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236, 2024.

[39] C. Song and Y. Song, "Enhancing academic writing skills and motivation: assessing the efficacy of chatgpt in ai-assisted language learning for efl students," *Frontiers in psychology*, vol. 14, p. 1260843, 2023.

[40] L. Wang and B. Ren, "Enhancing academic writing in a linguistics course with generative ai: An empirical study in a higher education institution in hong kong," *Education Sciences*, vol. 14, no. 12, p. 1329, 2024.

[41] J. Han, H. Yoo, Y. Kim, J. Myung, M. Kim, H. Lim, J. Kim, T. Y. Lee, H. Hong, S.-Y. Ahn, *et al.*, "Recipe: How to integrate chatgpt into efl writing education," in *Proceedings of the tenth ACM conference on learning@ scale*, pp. 416–420, 2023.

[42] P. Qin, C.-L. Yang, J. Li, J. Wen, and Y.-C. Lee, "Timing matters: How using llms at different timings influences writers' perceptions and ideation outcomes in ai-assisted ideation," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2025.

[43] O. Shaer, A. Cooper, O. Mokryn, A. L. Kun, and H. Ben Shoshan, "Ai-augmented brainwriting: Investigating the use of llms in group ideation," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024.

[44] Q. Wan, S. Hu, Y. Zhang, P. Wang, B. Wen, and Z. Lu, "" it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models," *Proceedings of the ACM on human-computer interaction*, vol. 8, no. CSCW1, pp. 1–26, 2024.

[45] Z. Wu, D. Ji, K. Yu, X. Zeng, D. Wu, and M. Shidujaman, "Ai creativity and the human-ai co-creation model," in *International Conference on Human-Computer Interaction*, pp. 171–190, Springer, 2021.

[46] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan, "Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice," *Computers in Human Behavior*, vol. 127, p. 107018, 2022.

[47] J. G. Boubin, C. F. Rusnock, and J. M. Bindewald, "Quantifying compliance and reliance trust behaviors to influence trust in human-automation teams," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, pp. 750–754, SAGE Publications Sage CA: Los Angeles, CA, 2017.

[48] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[49] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[50] D. Niraula, K. C. Cuneo, I. D. Dinov, B. D. Gonzalez, J. B. Jamaluddin, J. J. Jin, Y. Luo, M. M. Matuszak, R. K. Ten Haken, A. K. Bryant, *et al.*, "Intricacies of human–ai interaction in dynamic decision-making for precision oncology," *Nature Communications*, vol. 16, no. 1, p. 1138, 2025.

[51] G. Fragiadakis, C. Diou, G. Kousiouris, and M. Nikolaidou, "Evaluating human-ai collaboration: A review and methodological framework," *arXiv preprint arXiv:2407.19098*, 2024.

[52] N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes, "Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task," *arXiv preprint arXiv:2506.08872*, vol. 4, 2025.

[53] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review," *Smart Learning Environments*, vol. 11, no. 1, p. 28, 2024.

[54] C. Kobiella, T. Mitrevska, A. Schmidt, and F. Draxler, "When efficiency meets fulfillment: Understanding long-term llm integration in knowledge work," in *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, pp. 1–15, 2025.

[55] H. McNichols and A. Lan, "The studychat dataset: Student dialogues with chatgpt in an artificial intelligence course," *arXiv preprint arXiv:2503.07928*, 2025.

[56] J. He, S. Houde, and J. D. Weisz, "Which contributions deserve credit? perceptions of attribution in human-ai co-creation," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2025.

[57] M. Reza, J. Thomas-Mitchell, P. Dushniku, N. Laundry, J. J. Williams, and A. Kuzminykh, "Co-writing with ai, on human terms: Aligning research with user demands across the writing process," *arXiv preprint arXiv:2504.12488*, 2025.

[58] A. T. Wasi, M. R. Islam, and R. Islam, "Llms as writing assistants: Exploring perspectives on sense of ownership and reasoning," in *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pp. 38–42, 2024.

[59] J. J. Van der Bijl and L. M. Shortridge-Baggett, "The theory and measurement of the self-efficacy construct," *Self-efficacy in nursing: Research and measurement perspectives*, vol. 15, no. 3, pp. 189–20, 2002.

[60] S. Xu, K. I. Khan, and M. F. Shahzad, "Examining the influence of technological self-efficacy, perceived trust, security, and electronic word of mouth on ict usage in the education sector," *Scientific reports*, vol. 14, no. 1, p. 16196, 2024.

[61] L. Flower and J. R. Hayes, "The cognition of discovery: Defining a rhetorical problem," *College Composition & Communication*, vol. 31, no. 1, pp. 21–32, 1980.

[62] M. N. Siddiqui, R. D. Pea, and H. Subramonyam, "Script&shift: A layered interface paradigm for integrating content development and rhetorical strategy with llm writing assistants," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2025.

[63] C. W. VanVoorhis, B. L. Morgan, *et al.*, "Understanding power and rules of thumb for determining sample sizes," *Tutorials in quantitative methods for psychology*, vol. 3, no. 2, pp. 43–50, 2007.

[64] M. N. Siddiqui, V. Feliciano, R. Pea, and H. Subramonyam, "Ai in the writing process: How purposeful ai support fosters student writing," in *International Conference on Artificial Intelligence in Education*, pp. 190–203, Springer, 2025.

[65] A. Parabiaghi, H. E. Kortrijk, and C. L. Mulder, "Defining multiple criteria for meaningful outcome in routine outcome measurement using the health of the nation outcome scales," *Social psychiatry and psychiatric epidemiology*, vol. 49, no. 2, pp. 291–305, 2014.

[66] E. de Beurs, M. Blankers, P. Delespaul, E. van Duijn, N. Mulder, A. Nugter, W. Swildens, B. G. Tiemens, J. Theunissen, A. F. van Voorst, *et al.*, "Treatment results for severe psychiatric illness: which method is best suited to denote the outcome of mental health care?," *BMC psychiatry*, vol. 18, no. 1, p. 225, 2018.

[67] S. M. Haley and M. A. Fragala-Pinkham, "Interpreting change scores of tests and measures used in physical therapy," *Physical therapy*, vol. 86, no. 5, pp. 735–743, 2006.

[68] N. S. Jacobson and P. Truax, "Clinical significance: a statistical approach to defining meaningful change in psychotherapy research.," 1992.

[69] M. E. Abrahamse, M. Junger, E. L. Chavannes, F. J. Coelman, F. Boer, and R. J. Lindauer, "Parent–child interaction therapy for preschool children with disruptive behaviour problems in the netherlands," *Child and adolescent psychiatry and mental health*, vol. 6, no. 1, p. 24, 2012.

[70] T. Chanpradit, P. Samran, S. Saengpinit, and P. Subkasin, "English paraphrasing strategies and levels of proficiency of an ai-generated quillbot and paraphrasing tool: Case study of scientific research abstracts.," *Journal of English Teaching*, vol. 10, no. 2, pp. 110–126, 2024.

[71] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[72] G. Yaar and H. Erel, "Who's better? exploring the effects of social comparison in hri on self-esteem and self-efficacy," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 978–988, IEEE, 2025.

[73] R. Wiczorek and J. Meyer, "Effects of trust, self-confidence, and feedback on the use of decision automation," *Frontiers in psychology*, vol. 10, p. 519, 2019.

[74] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, and L. Kaplan, "Rapid trust calibration through interpretable and uncertainty-aware ai," *Patterns*, vol. 1, no. 4, 2020.

[75] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pp. 1–12, 2019.

[76] A. L. Alter and D. M. Oppenheimer, "Uniting the tribes of fluency to form a metacognitive nation," *Personality and social psychology review*, vol. 13, no. 3, pp. 219–235, 2009.

[77] R. Reber and C. Unkelbach, "The epistemic status of processing fluency as source for judgments of truth," *Review of philosophy and psychology*, vol. 1, no. 4, pp. 563–581, 2010.

[78] C. M. Griffiths, L. Murdock-Perriera, and J. L. Eberhardt, "“can you tell me more about this?”: Agentic written feedback, teacher expectations, and student learning," *Contemporary Educational Psychology*, vol. 73, p. 102145, 2023.

[79] D. Busby and C. Malone, "Writing circles: Developing learner self-efficacy and agency through peer review activities.," *Journal of Learning Development in Higher Education*, 2023.

[80] G. Pittam, J. Elander, J. Lusher, P. Fox, and N. Payne, "Student beliefs and attitudes about authorial identity in academic writing," *Studies in Higher education*, vol. 34, no. 2, pp. 153–170, 2009.

[81] M. Maguire, A. E. Reynolds, and B. Delahunt, "Self-efficacy in academic reading and writing, authorial identity and learning strategies in first-year students.," *All Ireland Journal of Higher Education*, vol. 5, no. 1, 2013.

[82] A. Fatima, K. K. Sunguh, A. Abbas, A. Mannan, and S. Hosseini, "Impact of pressure, self-efficacy, and self-competency on students' plagiarism in higher education," *Accountability in Research*, vol. 27, no. 1, pp. 32–48, 2020.

[83] A. Bodonhelyi, E. Bozkir, S. Yang, E. Kasneci, and G. Kasneci, "User intent recognition and satisfaction with large language models: A user study with chatgpt," *arXiv preprint arXiv:2402.02136*, 2024.

[84] M. Leijten and L. Van Waes, "Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes," *Written Communication*, vol. 30, no. 3, pp. 358–392, 2013.

[85] L. Tankelevitch, V. Kewenig, A. Simkute, A. E. Scott, A. Sarkar, A. Sellen, and S. Rintel, "The metacognitive demands and opportunities of generative ai," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

[86] M. N. Hoque, T. Mashiat, B. Ghai, C. D. Shelton, F. Chevalier, K. Kraus, and N. Elmqvist, "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*, pp. 1–15, 2024.

[87] B. Yuan and J. Hu, "Generative ai as a tool for enhancing reflective learning in students," *arXiv preprint arXiv:2412.02603*, 2024.

[88] M. Koskela, P. Luukkonen, T. Ruotsalo, M. Sjöberg, and P. Floréen, "Proactive information retrieval by capturing search intent from primary task context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 3, pp. 1–25, 2018.

[89] O. Press, A. Hochlehnert, A. Prabhu, V. Udandarao, O. Press, and M. Bethge, "Citeme: Can language models accurately cite scientific claims?," *Advances in Neural Information Processing Systems*, vol. 37, pp. 7847–7877, 2024.

[90] Y. Li, X. Yue, Z. Liao, and H. Sun, "Attributionbench: How hard is automatic attribution evaluation?," *arXiv preprint arXiv:2402.15089*, 2024.

[91] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–14, 2019.

[92] S. S. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, and O. Russakovsky, "Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2025.

[93] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: story writing with large language models," in *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 841–852, 2022.

[94] M. Urban, F. Děchtěrenko, J. Lukavskỳ, V. Hrabalová, F. Svacha, C. Brom, and K. Urban, "Chatgpt improves creative problem-solving performance in university students: An experimental study," *Computers & Education*, vol. 215, p. 105031, 2024.

[95] W.-F. Wang, C.-T. Lu, N. Ponsa i Campanyà, B.-Y. Chen, and M. Y. Chen, "Aideation: Designing a human-ai collaborative ideation system for concept designers," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2025.

[96] M. Zhang, H. Guo, and X. Liu, "Using keystroke analytics to understand cognitive processes during writing.," *International Educational Data Mining Society*, 2021.

[97] L. Shi, N. Bhattacharya, A. Das, M. Lease, and J. Gwizdka, "The effects of interactive ai design on user behavior: An eye-tracking study of fact-checking covid-19 claims," in *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pp. 315–320, 2022.

[98] Z. Wu, Y. Wang, M. Langer, and A. M. Feit, "Releyeance: Gaze-based assessment of users' ai-reliance at run-time," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 3, pp. 1–18, 2025.

[99] C. Qian and J. Wexler, "Take it, leave it, or fix it: Measuring productivity and trust in human-ai collaboration," in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 370–384, 2024.

[100] K. Zhou, J. D. Hwang, X. Ren, N. Dziri, D. Jurafsky, and M. Sap, "Rel-ai: An interaction-centered approach to measuring human-lm reliance," *arXiv preprint arXiv:2407.07950*, 2024.

[101] S. S. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan, "" i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust," in *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 822–835, 2024.

[102] A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou, and S. Koyejo, "Syceval: Evaluating llm sycophancy," *arXiv preprint arXiv:2502.08177*, 2025.

[103] H. Yoo, J. Han, S.-Y. Ahn, and A. Oh, "Dress: dataset for rubric-based essay scoring on efl writing," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13439–13454, 2025.

Acknowledgments in Korean

먼저 제가 연구자로 성장할 수 있도록 아낌없는 지도와 조언으로 이끌어주신 김주호 교수님께 감사드립니다. 항상 연구에 대해 함께 고민하고 유익한 피드백과 새로운 영감을 주신 KIXLAB 구성원들에게도 감사드립니다. 특히 졸업 연구를 함께 해주신 Nadia 님과 석사 과정 동안 훌륭한 멘토가 되어주신 김서영 님께 특별한 감사를 전합니다. 힘든 순간마다 제 고민을 들어주고 따뜻한 응원을 보내준 친구들에게도 고마운 마음을 전합니다. 특히 좌절할 때마다 나보다도 나를 믿어준 사람에게 감사합니다. 마지막으로 저를 격려하고 지원해주신 부모님과 가족들, 그리고 변함없이 제 편에서 힘이 되어준 동생에게도 깊은 감사를 드립니다.

Curriculum Vitae in Korean

이 름: 박 연 수
생 년 월 일: 2000년 09월 28일
전 자 주 소: yeonsupark@kaist.ac.kr

학 력

2019. 9. – 2024. 2. 한국과학기술원 전산학부 (학사)
2025. 3. – 2026. 2. 한국과학기술원 전산학부 (석사)

경 력

2024. 9. – 2025. 12. 한국과학기술원 전산학부 조교 (Introduction to Programming)
2024. 3. – 2024. 6. 한국과학기술원 전산학부 조교 (Introduction to Algorithms)

연 구 업 적

1. Jeongeon Park, Eun-Young Ko, **Yeon Su Park**, Jinyeong Yim, and Juho Kim. 2024. DynamicLabels: Supporting Informed Construction of Machine Learning Label Sets with Crowd Feedback. In Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24).
2. Hyoungwook Jin, Yoonsu Kim, **Yeon Su Park**, Bekzat Tilekbay, Jinho Son, and Juho Kim. 2024. Using Large Language Models To Diagnose Math Problem-solving Skills At Scale. In Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24).
3. Seoyoung Kim, **Yeon Su Park***, Dakyeom Ahn*, Jin Myung Kwak, and Juho Kim. 2024. Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker's Accent. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 169 (April 2024), 22 pages.
4. Juhoon Lee, Seoyoung Kim, **Yeon Su Park**, Juho Kim, Jeong-woo Jang, and Joseph Seering. 2025. Less Talk, More Trust: Understanding Players' In-game Assessment of Communication Processes in League of Legends. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).
5. Junho Myung*, **Yeon Su Park***, Sunwoo Kim*, Shin Yoo, and Alice Oh. 2025. PapersPlease: A Benchmark for Evaluating Motivational Values of Large Language Models Based on ERG Theory. In Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2), pages 522–531, Vienna, Austria and virtual meeting. Association for Computational Linguistics.