sustainability

IMPACT FACTOR 3.3

CITESCORE 7.7

Article

Prompt Engineering and Multimodal Tasks in AI-Supported EFL Education: A Mixed Methods Study

Debopriyo Roy, George F. Fragulis and Adya Surbhi

Special Issue

AI for Sustainable and Creative Learning in Education

Edited by

Dr. Chunfang Zhou, Prof. Dr. Connie Svabo and Dr. Serhii Petrovych

MDPI

https://doi.org/10.3390/su18052415

sustainability

MDPI

Article

Prompt Engineering and Multimodal Tasks in AI-Supported EFL Education: A Mixed Methods Study

Debopriyo Roy [1,*], George F. Fragulis [2] and Adya Surbhi [3]

[1] Technical Communication Laboratory, Center for Language Research, Department of Computer Science & Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Fukushima, Japan
[2] Department of Electrical and Computer Engineering, University of Western Macedonia, 50100 Kozani, Greece; gfragulis@uowm.gr
[3] Jindal Global Law School, O. P. Jindal Global University, Sonipat 131001, India; asurbhi@jgu.edu.in
* Correspondence: droy@u-aizu.ac.jp

Abstract

The rapid integration of artificial intelligence (AI) into higher education is reshaping how learners develop academic, linguistic, and research competencies. This mixed-methods study examines how second-year EFL computer science students employ prompt engineering techniques across four task domains—research summarization, academic video note-taking, style transformation, and concept mapping—within a smart learning environment. Sixty-nine students completed a structured survey requiring AI-assisted draft generation followed by student-led revision. Quantitative analyses included descriptive statistics, chi-square tests, Cramer's V, *t*-tests, ANOVA, Kruskal–Wallis tests, and three text-similarity measures (cosine, Jaccard, and Levenshtein). Qualitative evidence was drawn from students' revised outputs and reflective responses. Results indicate that students consistently preserved semantic meaning while significantly rephrasing AI-generated text, demonstrating moderate conceptual alignment but substantial lexical and structural transformation. Frequent AI users said they were better at searching and revising, but the type of prompt didn't have much of an effect on how deep the revision was or how well they learned. Iterative prompting and revision emerged as central drivers of metacognitive growth, academic language development, and sustainable learning behaviors. Across tasks, students viewed AI prompts as effective scaffolds for organizing information and synthesizing multimodal input, though reliance varied by learner. The findings underscore that sustainable AI use in EFL technical education depends not on AI output alone, but on structured prompting, iterative human revision, and critical engagement—practices that cultivate autonomy, digital literacy, and long-term academic resilience.

Keywords: artificial intelligence; prompt engineering; sustainable education; smart learning environments

Academic Editor: Fernando Moreira

Received: 9 December 2025
Revised: 9 January 2026
Accepted: 5 February 2026
Published: 2 March 2026

1. Introduction

Sustainability in education with AI means using AI in a way that is ethical and supports long-term, fair, and scalable learning systems that give people more control over their learning instead of making them dependent on it [1–3]. Sustainability in language education focuses on long-term learner growth, ecological thinking, and the creation of supportive learning environments instead of short-term performance improvements [4–6].

Sustainability in language education with AI means using AI tools in a long-term, ethical, and pedagogically sound way to help students become more independent, have more

control over their learning, and follow their own developmental paths. It also means keeping the learning environment in balance with nature, not relying too much on technology, and making sure that everyone has access to fair, human-centered learning opportunities. Sustainable practice emphasizes that AI should augment—not replace—meaningful social interaction, reflective learning, and identity development [6–9].

The rapid evolution of artificial intelligence (AI) is reshaping the foundations of technical education, particularly in domains such as computer science, where content mastery must be coupled with strong communication skills. AI tools, like smart tutoring systems and chatbots, are now essential in modern learning settings, providing personalized help, ongoing feedback, and flexible learning paths that support sustainable education. Evidence from systematic reviews indicates that AI-powered intelligent tutoring systems can support sustainable educational goals by promoting learner autonomy, adaptability, and long-term engagement when they are thoughtfully integrated into curricula [3].

The incorporation of Artificial Intelligence (AI) into education is transforming pedagogical frameworks, especially in technical fields like computer science. AI-driven tools, including intelligent tutoring systems and large language models, offer personalized learning experiences, adaptive feedback, and scalable support, aligning with the principles of sustainable education. These technologies enhance content mastery and foster critical thinking, problem-solving, and digital literacy among students [10].

In this changing world, prompt engineering—the careful design and improvement of instructions given to generative AI—has become a useful skill for both students and teachers. Scholars argue that prompt engineering functions as a new 21st-century digital literacy: it enables users to elicit more accurate, contextualized, and pedagogically useful outputs from generative models, and it can be taught and scaffolded within higher education curricula [11].

In the context of English as a Foreign Language (EFL) instruction, especially for students in technical fields, AI tools can bridge language barriers and facilitate the acquisition of academic English. However, the effectiveness of AI-generated content depends on students' ability to engage with and critically refine these outputs. This process necessitates the development of prompt engineering skills—crafting precise and contextually appropriate instructions for AI systems to generate desired responses. Research indicates that well-designed prompts can significantly improve the relevance and quality of AI-generated content, thereby enhancing learning outcomes [12].

Looking at prompt engineering in sustainable education in smart environments changes the view of technology from being a quick fix for learning to being a support that helps build lasting skills like critical thinking, understanding information, and turning machine-generated content into suitable language for their field. When learners actively revise AI outputs—rather than simply accepting them—they practice language editing, disciplinary reasoning, and ethical use of generative tools, all of which contribute towards incremental, constructive-deconstructive, and iterative processes of language acquisition. These strategies contribute to creating resilient learning ecosystems that align with sustainability objectives. This study elucidates how the implementation of structured tasks with prompts, coupled with guided revisions and a teacher's evaluation guide, can facilitate the development of a novel pedagogical approach that is adaptable, student-centered, and appropriate for contemporary educational settings.

Sustainable English education for non-native speakers needs teaching methods that help students improve their skills over time and ensure everyone has fair access to good learning materials. Recent studies highlight that sustainable EFL instruction must integrate emerging technologies in ways that strengthen all four language skills—reading, listening, writing, and speaking—while promoting learner autonomy and reducing in-

structional burden [13,14]. In this context, tools like ChatGPT 4.0 should not be used to create polished final products. Instead, they should be used as research assistants to help students go through cycles of inquiry, experimentation, and refinement. Multi-shot prompt engineering—where students work with several examples, guided prompts, and step-by-step changes—creates a better learning environment by helping them practice good reasoning, assess AI-generated content, and gradually improve their writing and analysis skills. This iterative engagement enhances language learning and deepens the research process itself, fostering metacognitive awareness, critical evaluation, and more sophisticated academic outcomes overall [15,16]. As large language models play a bigger role in how we use language and do knowledge work, including prompt engineering in effective teaching methods provides a way to build strong communication and research skills for the future.

This study looks at how EFL computer science students used prompt engineering techniques in the four language skills—reading, listening, writing, and speaking—by including specific AI-related tasks (like summarizing articles, taking notes from lectures, creating concept maps, scripting presentations, and designing posters) in their course. By analyzing assignments such as summarizing scholarly articles, taking structured notes from videos, constructing concept maps, preparing technical presentations, and designing research posters, the research aims to assess students' ability to utilize AI outputs effectively and revise them to meet academic standards. The results aim to guide educational approaches that incorporate AI tools in a way that fosters enduring learning habits and equips students for forthcoming challenges in a technology-centric society. To understand both how students act and what they learn, the survey gathered numerical data on how they use prompts and revise their work and combined this with examples of AI-generated text and the students' improved versions. Previous empirical work suggests that such an approach (teaching students to craft prompts and then critically revise AI outputs) can enhance metacognitive awareness and language production in EFL contexts [16].

In summary, as AI continues to shape the educational landscape, it is imperative to equip students with the skills to navigate and leverage these technologies effectively. By concentrating on prompt engineering and critical interaction with AI-generated content, educators can cultivate a learning environment that is not only technologically sophisticated but also sustainable, inclusive, and responsive to the changing requirements of the digital era.

2. Significance of the Study

This research contributes to the evolving landscape of sustainable education by exploring how AI-driven tools, specifically prompt engineering, can enhance students' research capabilities in technical disciplines. Incorporating AI into education aligns with the principles of sustainable development by promoting personalized learning, fostering critical thinking, and preparing students for the demands of the digital age. As highlighted by [10], AI-driven adaptive learning technologies are transforming education by making it more personalized and accessible, which is crucial for sustainable educational practices.

Furthermore, integrating prompt engineering into the research process empowers students to interact more effectively with AI systems, enhancing their ability to identify, evaluate, and utilize scholarly resources. This skill is increasingly vital in a world where information is abundant, but the ability to discern and apply relevant knowledge is paramount. Ref. [3] discusses how AI in intelligent tutoring systems can improve access to quality education, creating personalized learning experiences that are essential for sustainable education.

By examining the intersection of AI tools and research practices, this study offers ideas about how educational technologies can be leveraged to meet real-world requirements. The findings emphasize the value of equipping students with the skills to navigate and utilize AI-driven resources effectively, thereby enhancing their research competencies and preparing them for future challenges in a technology-driven world.

3. A Review of the Literature

Scholarly attention has grown toward the role of prompt engineering in language education, particularly in EFL settings. In human–AI collaborative story-writing tasks, a study of 67 Hong Kong secondary school students revealed that learners' prompting practices evolved: they used prompts to overcome writer's block, to develop and expand narrative ideas, and to improve the coherence of their stories [17,18]. This study—based on activity theory—further suggests that students' awareness of the purpose of prompting is critical: when they understand why they are issuing a prompt (e.g., for brainstorming, refining, or translating), they can scaffold more effectively and shape more meaningful AI-generated texts [18].

Another empirical investigation traced how secondary-level EFL students approached ChatGPT 3.5 in a writing task, revealing distinct "prompt-engineering pathways" characterized by trial-and-error, differing in both the quantity and quality of prompts used [19]. This kind of reflective practice—iteratively refining prompts, observing outcomes, and adapting—supports not only higher-quality AI output but also learners' metalinguistic awareness and writing competence.

Prompt engineering also supports academic English development. Recent research has shown that prompt engineering significantly boosted students' self-regulated learning (SRL) behaviors and improved their proficiency in academic English (for non-native speakers) through the use of an LLM-powered learning tool [20].

From the applied linguistics perspective, scholars have articulated explicit strategies for prompt design, breaking down its key components (e.g., persona, audience, context, instructions, and output specification) and proposing techniques such as iterative prompting, role-prompting, and few-shot prompting [21]. Such frameworks act like support structures that help both learners and teachers to enhance the quality of prompts, minimize errors or biases from AI, and better align the results with language learning objectives.

Practically, prompt engineering has been shown to improve EFL writing outcomes. In a recent conference study involving Japanese undergraduates, instructors demonstrated how carefully designed prompts enhanced ChatGPT's effectiveness in supporting academic writing tasks while also identifying strategies for mitigating issues related to academic integrity [22]. Meanwhile, teacher-focused research echoes the need for prompt literacy: in a study of ESL/EFL educators, teachers reported that prompt-engineering techniques enabled them to generate a wider variety of task-appropriate teaching materials (e.g., quizzes, dialogues, and role-plays) using LLMs, though they also raised concerns about accuracy and factual reliability—underscoring the importance of critical evaluation [23].

3.1. Adoption and Use of AI in Education

Recent evidence points to a dramatic increase in student adoption of generative AI tools across higher education contexts. A large-scale survey in the United Kingdom reported that 92% of university students have used tools such as ChatGPT 4.0, a significant rise from 66% recorded the previous year [24,25]. This surge highlights a shifting academic landscape where AI usage is quickly becoming a normalized part of students' everyday learning practices. Despite this rapid uptake among learners, faculty readiness remains uneven. For example, in Indian business schools, only 7% of faculty members identify them-

selves as expert AI users, pointing to a substantial professional development gap [26,27]. This discrepancy points to the need for targeted training to align faculty competencies with student expectations and contemporary educational demands.

3.2. AI's Role in Research and Literature Review

Both students and educators are increasingly adopting AI-assisted research practices, which are revolutionizing the way they conduct literature searches and academic inquiries. A recent study revealed that 51% of students and researchers now rely on AI tools to support literature reviews, improving speed and depth of information retrieval [28]. Similarly, AI has become embedded in the everyday workflow of educators, with 44% of teachers incorporating AI for tasks such as resource gathering, research support, and instructional preparation [29]. These patterns demonstrate how AI is reshaping academic research by streamlining cognitive load and facilitating more efficient access to scholarly information.

3.3. Prompt Engineering in Education

Prompt engineering has become an essential skill in modern educational settings, with increasing evidence illustrating its influence on critical thinking, analytical reasoning, and individualized learning experiences.

Research shows that well-designed prompts in AI-enhanced learning environments foster higher-order thinking skills by supporting student reasoning and reflection. Meta-analytic evidence suggests that prompts that encourage reflection, breaking down problems, and reasoning help students think critically and at a higher level in digital learning environments [30]. Socratic-style chatbot prompts encourage students to think critically and reflect on their work, which helps them solve problems and think strategically more effectively than regular prompts [31]. Similarly, systematic reviews highlight that prompt engineering in higher education can scaffold complex learning tasks and align questioning strategies with educational goals [32]. Empirical studies demonstrate that AI chatbots with adaptive prompts support higher-order thinking across multiple cognitive domains, particularly in writing and problem-solving tasks [33,34]. Techniques like least-to-most prompting, which breaks down hard problems into smaller ones, help students think more clearly and get more involved in higher-order cognitive processes [35]. Overall, this body of research points out the importance of thoughtfully designed prompts in promoting analytical, evaluative, and reflective thinking in AI-supported learning contexts [30–35]. Collectively, these findings highlight prompt engineering not merely as a technical skill but as a key literacy for navigating AI-mediated knowledge ecosystems.

3.4. Sustainable Education and Real-World Applications

Governments and educational institutions are increasingly integrating AI literacy into formal curricula to ensure readiness for future societal and workforce demands. For instance, the Punjab government in India has initiated a province-wide AI curriculum to align school instruction with global technology standards and promote sustainable digital education [27]. Beyond curriculum reform, the broader societal expectation for AI competence is evident among learners themselves. A recent survey found that 73% of college students consider AI-related skills as the most essential part of their higher education experience, reflecting shifting priorities regarding career preparedness and technological fluency [36]. This trend underscores the urgency for education systems to embed AI competencies, such as prompt engineering, into mainstream learning pathways.

3.5. Thematic Integration and Sustainability-Oriented Research Gap

Recent studies on AI-assisted language learning have shown that large language models can improve writing fluency, idea generation, and basic accuracy in English as a

Foreign Language (EFL) settings. However, much of this work conceptualizes effectiveness in terms of immediate performance gains or output quality, often treating AI tools as productivity enhancers rather than as components of a sustainable learning ecology. As a result, key questions remain unresolved regarding how learners interact with AI over time, how agency and critical engagement are maintained, and how AI-mediated practices align with long-term educational sustainability.

From a sustainability viewpoint, current studies often focus on access and efficiency but do not adequately address important aspects like learner independence, mental involvement, and ethical use—key parts of sustainable education models. Based on learning theories that focus on sustainability, the literature suggests that for AI to be used effectively in education, students need to actively assess, change, and relate to AI results instead of just accepting them without question. Yet empirical evidence detailing *how* such processes unfold in authentic classroom settings—particularly in technical EFL disciplines—remains limited.

Contribution—The present study addresses these gaps by shifting the analytical focus from AI output quality to *revision depth, iterative prompting, and meaning preservation* as indicators of sustainable learning practice. Rather than asking whether AI improves writing, this study examines how students negotiate control, responsibility, and discourse norms while using AI tools. By using text similarity analysis, statistical models, and student reflections, the study defines sustainability as finding a balance between help from technology and human decision-making. In this way, it directly answers calls in the literature for process-based, ethically sound, and learner-centered studies of AI-mediated education. It positions prompt engineering not as a quick fix, but as a long-term way to improve academic literacy.

4. Research Questions

These typical research questions will emerge within the context of the current study. These questions will be answered in a limited way in the context of this study, which should be considered more of a case study in an EFL context for a computer science university.

4.1. RQ1: Use of AI Prompts

- How frequently do EFL computer science students use AI prompt engineering techniques across different language skill tasks (reading, listening, speaking, and writing/visual)?

In their research, [13] demonstrates the frequent and varied applications of AI in reading, writing, and vocabulary tasks for EFL learners.

- Which types of prompts, such as summarization, note extraction, concept mapping, presentation scripting, and poster drafting, are students most commonly using?

Ref. [14], in their research, identified common prompt types, including summarization, outlining, explanation, and multimodal task prompts.

4.2. RQ2: Revision Practices

- To what extent do students revise AI-generated outputs before submission?

  Ref. [15] examined revision extent and revision depth when students use AI drafts.

- What strategies do students use to revise AI-generated text (e.g., rephrasing, simplifying technical terms, restructuring ideas)?

In their research, ref. [37] emphasized document strategies including rephrasing, simplifying, reorganizing concepts, and rectifying discipline-specific terminology.

- How does the degree of revision vary across different assignment types (e.g., summaries, notes, concept maps, presentations, and posters)?

Revision plays a critical role in the context of AI-prompted writing because it enables more profound engagement with generated text beyond surface corrections. For example, ref. [38] found in their study of L2 writers that even after a month of focused academic writing lessons, the changes in how they revised their work were small—indicating that students don't automatically make deeper changes to their writing without help. Similarly, ref. [39] shows that when students receive rubric-plus-exemplar feedback, they make more deep-level revisions in the structure and coherence of argumentative essays than when feedback is just in-text comments. In AI-assisted writing, this insight implies that merely generating text via prompts is insufficient; learners (or users) also need structured feedback and revision scaffolding to transform initial drafts into higher-quality, meaningful outputs.

4.3. RQ3: Impact on Learning Outcomes

- How does the use of prompt engineering influence students' development of English language skills in a technical context?

Ref. [40], in their research, demonstrated improved technical-register proficiency through structured prompting.

- How does AI-assisted revision affect the clarity, accuracy, and disciplinary appropriateness of student submissions?

Ref. [41], with their intriguing work, demonstrated improvements in clarity and accuracy when students revise AI-generated drafts.

- Does engagement with AI prompts enhance students' critical thinking, metacognitive awareness, and ability to synthesize technical content?

Ref. [42] reported gains in metacognitive monitoring and critical synthesis through reflective prompting.

4.4. RQ4: Task-Specific Effectiveness

- Which assignments (reading, listening, integrated concept mapping, presentation, and poster design) benefit most from AI prompt use and revision?

Ref. [43] showed differential benefits for summarization, concept mapping, and oral presentation tasks.

- How do students perceive the usefulness of AI-generated outputs for different tasks?

Ref. [44] reported varying perceptions depending on task complexity and modality.

4.5. RQ5: Sustainability and Smart Learning

- How does prompt engineering help smart learning environments use educational methods that are beneficial for the environment?

Ref. [45] showed efficient, scalable learning support through structured AI interactions.

- Can integrating AI-assisted tasks with structured revision promote long-term learner autonomy, digital literacy, and adaptive skills in technical education?

Recent research highlights the importance of AI literacy and learner autonomy in higher education. Ref. [46] asserts that students' AI literacy is influenced by their self-regulated learning strategies, indicating that addressing learners' needs improves their capacity to effectively utilize AI tools. In the same way, ref. [47] suggests a framework for AI literacy that shows different groups of learners need specific approaches to build their AI skills, which helps them learn independently and effectively in education that uses AI.

4.6. RQ6: Challenges and Limitations

- What challenges do students face when using AI prompts (e.g., vague outputs, over-reliance on AI, difficulty in revision)?

  Ref. [48] detailed common issues and learner struggles.

- How do these challenges differ across language skills and assignment types?

  Ref. [49] showed challenge patterns vary by task modality and cognitive load.

The above research questions will be attempted as part of the four sections requiring AI-assisted draft generation, followed by student-led revision.

4.7. Consolidated Overarching Research Questions

ORQ1: Learner–AI Interaction and Revision Processes

How do EFL computer science students engage with AI tools across academic tasks, and how do their prompt use and revision practices reflect depth of engagement, meaning preservation, and control over AI-generated text?

(Integrates original RQ1 and RQ2)

ORQ2: Task-Sensitive Learning Outcomes and Academic Literacy Development

How do AI-assisted prompting and revision practices shape students' academic language use, disciplinary appropriateness, synthesis abilities, and metacognitive awareness across different task types?

(Integrates original RQ3 and RQ4)

ORQ3: Sustainability, Agency, and AI Literacy

To what extent do AI-assisted tasks combined with structured revision support sustainable learning practices, including learner autonomy, critical evaluation of AI output, and the development of AI-related academic literacies in technical EFL contexts?

(Integrates original RQ5)

ORQ4: Constraints, Challenges, and Boundary Conditions

What challenges and limitations do students encounter when using AI tools for academic work, and how do these challenges vary by task demands and language skill focus?

(Integrates original RQ6)

5. Methods

5.1. Participants

The participants in this study consisted of second-year undergraduate computer science students enrolled in an English as a Foreign Language (EFL) 4-skill course at a Japanese computer science university. The sample was selected because these students represent a population situated at the intersection of technical education and language learning. Their coursework involves listening, reading, writing, and speaking activities with a focus on academic and technical communication, making them an ideal group for exploring the application of AI-driven prompt engineering in education. 69 students participated in the survey, providing a balance of perspectives while maintaining a manageable dataset. However, some participant data were excluded because they did not follow the assignment instructions carefully.

5.2. Sample

For this study, the sample consisted of 69 undergraduate computer science majors between the ages of 18 and 20, with male students comprising the majority. Although the participants regularly use ChatGPT 3.5 and other large language model (LLM) tools, they do not possess advanced proficiency in prompt engineering techniques in either Japanese or English. They are able to generate basic prompts without guidance; however, with

instructional support, they can produce simple follow-up prompts in English and more advanced prompts when they first formulate them in Japanese and translate them into English. Consequently, their overall ability to craft thoughtful, well-structured prompts remains limited.

5.3. Sampling Justification

Second-year students were chosen intentionally for this research. In their second year, computer science majors at Japanese universities usually have a basic understanding of both programming and English language courses, but they haven't yet focused on more advanced areas of computer science. This timing makes them a crucial group for testing how AI-assisted prompting and revision practices can support sustainable education goals, as they are still shaping their academic research habits and communication skills.

Japanese EFL learners frequently encounter difficulties in applying English for academic and field-specific purposes [50,51], and their experiences are shaped by the broader shift toward global English types in higher education [52]. Studies on AI in education suggest that carefully designed AI-based interventions, such as prompt engineering, can foster both language development and critical digital literacy [53–55]. The choice of this group reflects the broader educational need to equip future technical professionals with sustainable research and communication practices that align with real-world academic and industry requirements.

5.4. Construct Validity and Operationalization of Key Variables

To ensure construct validity, the study explicitly operationalizes revision depth and learning outcomes using measurable, theory-aligned indicators rather than relying on subjective or global judgments. Revision depth was defined as how much and in what way the AI-generated text was changed, measured using different methods that look at meaning, word choice, and surface-level similarities. Specifically, cosine similarity was used to assess semantic retention, the Jaccard Index to quantify lexical overlap, and Levenshtein similarity to capture character-level rewriting effort. Together, these measurements help to differentiate between simple edits and more significant changes in structure and wording, showing that revision depth reflects a real understanding of the content instead of just changing the text.

Learning outcomes were operationalized as process-oriented academic literacy gains rather than direct measures of content mastery. These outcomes were measured by looking at how well students could (a) switch between casual and academic writing styles, (b) keep the same meaning while changing the wording, (c) improve their prompts based on previous AI responses, and (d) carefully check and fix any style or factual mistakes made by the model. Self-reported perceptions of comprehension, efficiency, and task usefulness were analyzed in conjunction with observed revision patterns and similarity metrics to triangulate learning-related behaviors. The study therefore interprets learning outcomes as evidence of developing academic discourse competence and AI-mediated revision skills, not as long-term language acquisition or domain knowledge gains, which fall outside the scope of the present design.

5.5. Reproducibility and Qualitative Coding Constraints

The absence of formal inter-rater reliability measures inherently limits the reproducibility of the qualitative components of this study. Qualitative interpretations—especially those about learner reflections and revision behaviors—were done by just one researcher to keep the analysis consistent and in line with the study's theory. As a result, these interpretations should be understood as context-sensitive and illustrative rather than independently verifiable through coder agreement. To mitigate this limitation, qualitative observations were not

used as standalone evidence but were triangulated with objective text-similarity metrics and aggregated statistical patterns. This approach lessens the dependence on personal opinions for key analysis conclusions, while recognizing that future research using several coders and formal reliability tests would improve the consistency of qualitative analyses.

5.6. Instruments

The primary instrument used in this study is a structured survey in Google Forms, centered on EFL learning tasks, including note-taking from academic articles, summarizing texts, watching instructional videos, creating concept maps, designing technical presentations, and developing research posters. The four parts of the survey all include AI prompt engineering tasks. In these tasks, students must (a) use AI tools to come up with answers, (b) change or rephrase their answers to make them more appropriate for school, and (c) think about how they did it.

5.7. The Survey

This study was conducted in four structured sections designed to develop AI-assisted academic writing and research skills among computer science students.

5.7.1. AI-Supported Research and Summarization (Section 1 of the Survey)

Students selected a computer science topic relevant to their coursework and used Google Scholar to locate scholarly articles. Through prompt engineering (e.g., "Find the most cited articles on [topic]"), they refined searches, generated AI-based summaries, and evaluated article relevance. They then revised the AI-generated summaries in their own words before submitting both versions as evidence.

5.7.2. Academic Video Summarization (Section 2 of the Survey)

Students watched a scholarly computer science video and used AI prompts to produce a structured summary. They rephrased this output into academic English, demonstrating comprehension and synthesis of visual content for academic purposes.

5.7.3. Prompt Engineering for Style Transformation (Section 3 of the Survey)

Participants selected an informal paragraph and applied AI prompts to rewrite it into two academic versions using different prompting styles (e.g., role prompting, tone control). They then converted one academic version back into casual English, comparing tone, structure, and vocabulary shifts.

5.7.4. AI-Supported Concept Mapping (Section 4 of the Survey)

Students selected one article and one video on the same topic, used AI to generate comparative outlines, and transformed the AI text into concise notes for a visual concept map. This map illustrated thematic relationships using hierarchical and networked structures.

Figure 1 demonstrated an explanation of survey sections. To complete the survey, students reflected on which AI prompts were most effective, how they revised AI-generated text, the challenges faced in maintaining tone and meaning, and how AI-supported activities enhanced their understanding and organization of computer science concepts.

Figure 2 presents an ontology-styled diagram that visualizes the conceptual structure of the study and the relationships among its four main instructional and analytical sections. The diagram illustrates how each section is not treated as an isolated task but as part of an interconnected learning and analysis cycle in which outputs from one section inform subsequent activities. In particular, the Survey/Reflection node is positioned as a central feedback mechanism, capturing learners' perceptions, strategies, and challenges and feeding this information back into each section. This feedback loop highlights the iterative

nature of the design, showing how student reflection supports refinement of prompt use, revision behavior, and task engagement across sections, thereby reinforcing the study's process-oriented and sustainability-focused framework.

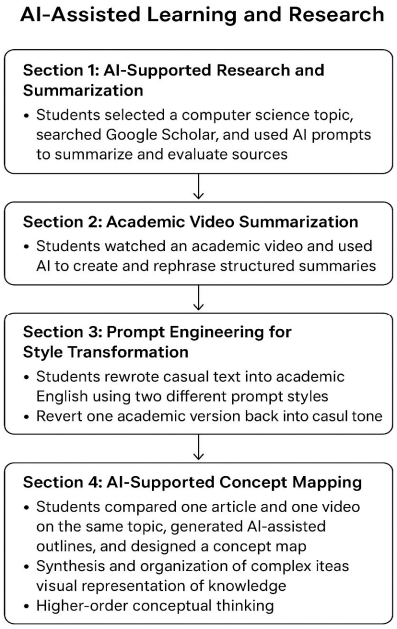


Figure 1. AI-assisted Learning and Research Model.

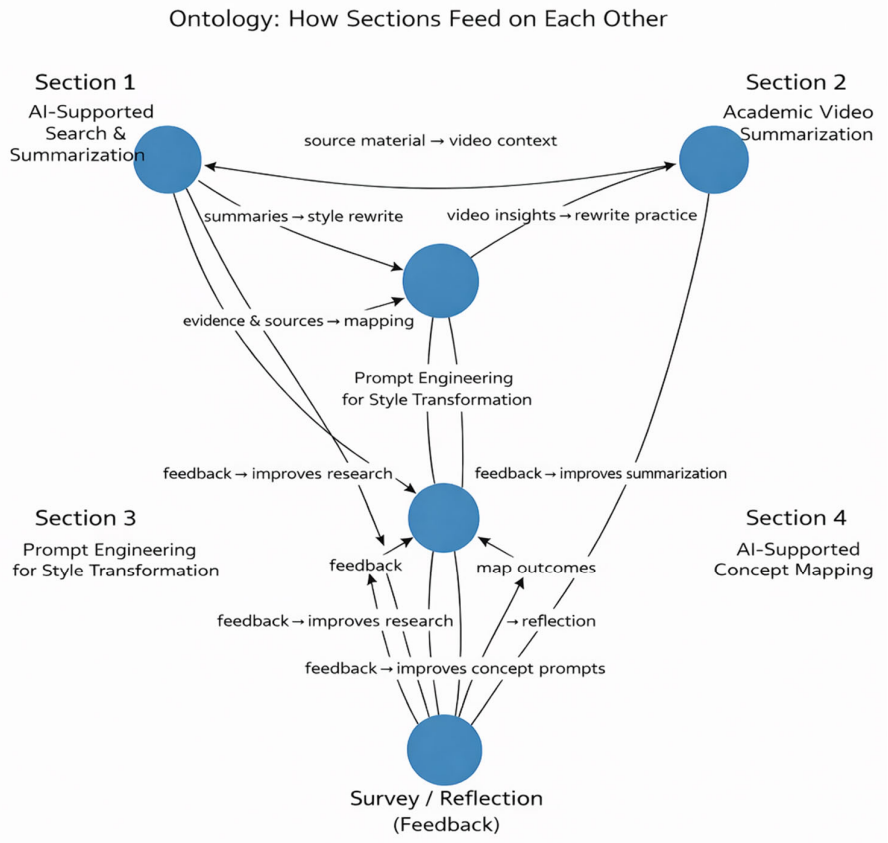


Figure 2. An Ontology-style Diagram.

Table 1 helps us understand the ontology. Table 2 is focused on the learning outcomes resulting from each section.

Table 1. Sections, Activities & Measures, and How They Connect to Other Sections.

Section	Main Activity	What It Measures/Develops	Connections to Other Sections
Section 1: AI-Supported Research and Summarization	Students selected a computer science topic, searched Google Scholar, and used AI prompts to summarize and evaluate sources.	• Research literacy • Ability to use AI for academic search and summarization • Critical evaluation of AI outputs	It lays the groundwork for subsequent tasks by honing skills in source analysis and summarization.
Section 2: Academic Video Summarization	Students watched an academic video and used AI to create and rephrase structured summaries.	• Listening and comprehension of scholarly content • Integration of AI in multimodal learning • Academic writing from non-text sources	The integration of AI in multimodal learning builds on Section 1 by applying summarization skills to visual content, preparing for the style transformation in Section 3.
Section 3: Prompt Engineering for Style Transformation	Students rewrote casual text into academic English (two styles) and reverted one academic version back into a casual tone.	• Prompt Engineering Skills • Control over tone, vocabulary, and structure • Understanding of academic vs. conversational style	Strengthens language transformation skills needed for clarity in concept mapping (Section 4).
Section 4: AI-Supported Concept Mapping	Students compared one article and one video on the same topic, generated AI-assisted outlines, and designed a concept map.	• Synthesis and Organization of Complex Ideas • Visual Representation of Knowledge • Higher-Order Conceptual Thinking	Integrated insights from Sections 1–3 to demonstrate full-cycle understanding of AI-supported learning and writing.

Table 2. Learning Outcomes Resulting from Each Section.

Section	Learning Outcomes
Section 1: AI-Supported Research and Summarization	Demonstrate research literacy through identifying credible academic sources. —Use AI tools effectively for academic search and summarization. —Critically evaluate AI-generated summaries and refine them for accuracy. —Develop foundational skills for multimodal and advanced tasks in later sections.
Section 2: Academic Video Summarization	Comprehend and extract key ideas from scholarly video content. —Summarize multimodal academic information using AI. —Apply summarization strategies learned in Section 1 to non-text sources. —Strengthen academic expression when converting visual/audio input into structured summaries.

Table 2. *Cont.*

Section	Learning Outcomes
Section 3: Prompt Engineering for Style Transformation	Apply prompt engineering strategies to transform text across styles and registers. —Develop control over academic tone, vocabulary, and structure. —Distinguish between academic and conversational writing styles. —Improve clarity and precision in language for later conceptual and analytical tasks.
Section 4: AI-Supported Concept Mapping	Synthesize ideas from multiple sources (video + article) into organized structures. —Create coherent visual representations of knowledge through AI-assisted mapping. —Demonstrate higher-order thinking by identifying relationships among concepts. —Integrate skills from previous sections to produce holistic, AI-supported academic outputs.

5.8. Procedure

The survey was given out in the second half of the semester, after students had already learned the basics of academic research and technical communication in English. In each section, participants were given clear instructions on how to use AI prompts to write first drafts or responses and then revise them to make them ready for submission.

To collect meaningful evidence, the survey required students to submit both AI-generated text and their revised version, enabling analysis of how AI assistance influenced their note-taking, summarizing, and presentation preparation processes. Multiple-choice items and open-ended reflection prompts captured students' perceptions of prompt effectiveness, their strategies for revision, and their confidence in using AI tools for academic work.

5.9. Data Availability Statement

The data generated and analyzed during this study include anonymized survey responses, AI-generated and student-revised text samples, and derived similarity metrics (cosine similarity, Jaccard index, and Levenshtein similarity), as well as aggregated statistical outputs. Due to ethical restrictions related to participant privacy and institutional consent agreements, the raw textual data and individual-level responses are not publicly available. However, anonymized aggregated data and statistical summaries supporting the findings of this study are available from the corresponding author upon reasonable request.

5.10. Data Analysis

The survey will be analyzed using a mixed-methods approach, combining quantitative and qualitative techniques to capture both measurable trends and more profound insights into students' use of AI prompts in EFL tasks. Table 3 has demonstrated the tests undertaken for each section and what it suggested.

5.11. Methodological Justification

This study adopts an exploratory mixed-methods design to examine how EFL computer science students engage with AI-generated text in authentic instructional settings. The methodological focus is on process and interaction rather than causal comparison, prioritizing ecological validity in a context where AI tools are already embedded in coursework.

Table 3. Master Table: Sections, Research Questions, Analyses, Procedures, and Purposes.

Section	Research Question (RQ)	Analysis Components Used	What Was Done (Procedures in Findings)	Purpose/What It Answered
1	RQ1: How do students revise? How does AI-generated text relate to revision and learning outcomes, such as comprehension and listening?	- Descriptive statistics - Chi-square tests - Cramer's V - *t*-tests - One-way ANOVA - Jaccard Similarity - Cosine Similarity (TF–IDF) - Levenshtein Similarity - Combined similarity interpretation - Survey–text triangulation	- Calculated revision levels and distribution across students - Ran Chi-square tests for prompt type × revision and revision × comprehension improvement - Computed effect sizes (Cramer's V) for all significant relationships - Used *t*-tests and ANOVA to compare groups based on revision depth - Measured lexical (Jaccard), semantic (Cosine), and structural (Levenshtein) similarity between AI-generated and student-revised text - Cross-checked students' self-reported revision behavior with actual similarity metrics	- Determined whether revision depth is associated with comprehension gains - Identified whether students revise AI-generated text superficially or deeply - Established how much of the AI-generated text students change (lexically, semantically, structurally) - Showed that prompt type does not dictate revision behavior - Validated that deeper revision is linked to higher comprehension
2	RQ2: How do students transform AI-generated academic video summaries, and what revision patterns influence comprehension?	- Word-count descriptive statistics - Chi-square & Cramer's V - Targeted *t*-tests - One-way ANOVA - Post hoc tests (if needed) - Similarity cross-validation	- Compared AI-generated vs. student-revised summary lengths - Ran Chi-square tests for revision × comprehension - Computed medium effect sizes for comprehension outcomes - Conducted ANOVA to compare revision patterns across video topics - Examined similarity metrics to confirm depth of revision	- Identified how students condense, refine, and reorganize AI summaries - Established that deeper revision is associated with higher comprehension - Showed whether topic type influences revision effort - Confirmed similarity scores mirror revision patterns
3	RQ3: How do students modify AI-generated text when shifting between academic and conversational styles?	- Open coding - Thematic coding - Correlation analysis (r) - One-way ANOVA - Tukey HSD post hoc tests - Jaccard Similarity - Cosine Similarity - Levenshtein Similarity - Strategy-to-text mapping	- Coded all student revisions to identify strategies (simplification, expansion, rephrasing, tone shift) - Grouped codes into themes (style adjustment, clarity improvement, structural reorganization) - Measured correlations between revision amount and similarity scores - Conducted ANOVA and post hoc tests to compare groups - Calculated three types of similarity to evaluate depth and type of transformation	- Identified how students interpret and execute style transformation tasks - Measured semantic vs. lexical vs. structural changes - Mapped methodological strategies to actual text edits - Showed how meaning is preserved but tone is adapted
4	RQ4: How do students evaluate the effectiveness of AI prompts when constructing concept maps, and do topic types influence these evaluations?	- Descriptive statistics - Chi-square tests - Cramer's V - Kruskal–Wallis test - Violin plot interpretation - Survey–theme cross-validation - Integrated interpretation	- Calculated effectiveness ratings for concept-map prompts - Ran Chi-square for topic × effectiveness and evaluated Cramer's V - Used Kruskal–Wallis to test non-parametric differences across topics - Interpreted violin plots to understand rating distribution - Connected qualitative comments to quantitative ratings	- Determined how students judge prompt usefulness in higher-order tasks - Evaluated whether topic type influenced perceived effectiveness - Identified convergence and divergence between numerical ratings and written explanations - Provided a holistic understanding of prompt utility in conceptual tasks

Text-similarity metrics (cosine, Jaccard, and Levenshtein) were used to quantify revision behavior across semantic, lexical, and structural dimensions. These measures do not assess writing quality directly but provide objective indicators of how extensively students transformed AI-generated text while preserving meaning. Inferential analyses (chi-square tests and ANOVA) were selected to examine task-level and prompt-related patterns and were interpreted conservatively within an exploratory framework.

A traditional non-AI control group was not included because the primary aim of the study was not to compare AI-assisted instruction with non-AI instruction but to examine how students interact with, revise, and critically evaluate AI-generated text within an authentic learning environment where AI use is already normalized. Removing or restricting AI access would have created artificial learning conditions that diverge from current instructional practice and limited ecological validity.

Moreover, the study focuses on *process-oriented outcomes*—such as revision depth, semantic preservation, and synthesis strategies—that are not meaningfully captured through between-group comparisons alone. Instead, the design relies on within-task and cross-task analyses, multiple text-similarity metrics, and learner reflections to reveal variation in engagement and revision behavior. This approach allows the study to identify meaningful learning patterns without positioning AI's use of itself as the experimental variable.

Finally, because large language models and AI policies are changing quickly, comparing a fixed control group may not provide clear results. The chosen design offers a practical way to understand new AI-related reading and writing skills while also setting the stage for future studies that might include control groups once teaching methods become more stable. The pedagogical approach and institutional policies about its interpretation still lack clarity about AI use.

Overall, the methodology balances analytic rigor with instructional realism and is appropriate for investigating emerging AI-supported literacies in EFL contexts.

5.12. Methodological Limitations

Several methodological limitations should be considered when interpreting the findings. First, the study is confined to a single institutional context and a relatively homogeneous sample of second-year computer science majors at a Japanese university. While this group is theoretically appropriate for examining AI-supported EFL instruction in technical education, the narrow context limits generalizability across disciplines, proficiency levels, and educational settings.

Second, the study is based on course-embedded assignments and survey data rather than an experimental design. The absence of random assignment and a non-AI control group limits causal inference; observed relationships between prompt engineering, revision behavior, and learning outcomes should therefore be interpreted as correlational.

Third, part of the data relies on students' self-reported perceptions of AI use and revision practices. Although these reports were triangulated with objective text-similarity measures (cosine, Jaccard, and Levenshtein), self-report bias and differences in scale interpretation may remain. Discrepancies between perceived and actual behavior remain possible.

Fourth, the text-similarity metrics capture lexical, structural, and semantic overlap between AI-generated and revised texts but do not fully reflect qualitative aspects of writing quality such as rhetorical effectiveness, disciplinary appropriateness, coherence, or originality. These measures should therefore be interpreted as indicators of revision depth rather than comprehensive assessments of learning or writing proficiency.

Fifth, individual differences in prior AI experience, English proficiency, and learning strategies were not controlled. Although participants shared similar academic backgrounds,

variation in familiarity with AI tools and metacognitive skills may have influenced prompting and revision behavior, potentially confounding some findings.

Sixth, the participant sample exhibited a gender imbalance that reflects enrollment patterns commonly observed in computer science programs within the study context. While this distribution is representative of the instructional setting, it may have implications for the generalizability of the findings. Prior research suggests that gender can influence technology engagement, help-seeking behavior, and attitudes toward AI-supported learning, which may in turn shape prompting strategies, revision practices, and perceptions of usefulness. As the present study did not examine gender-based differences, the results should be interpreted cautiously and not assumed to generalize uniformly across genders. Future research with more gender-balanced or explicitly comparative samples is needed to investigate whether learner–AI interaction patterns and sustainability-related outcomes differ by gender.

Finally, the results are shaped by the capabilities of AI tools available at the time of data collection. As large language models and institutional AI policies continue to evolve, the reproducibility and applicability of these findings may change. Accordingly, the study should be viewed as context-specific and exploratory rather than definitive.

6. Results

6.1. Section 1

Section 1 provides a thorough look at how participants responded to the task of finding research articles, showing how students found, understood, and sorted scholarly sources both with and without help from AI-based tools. This section illustrates the accuracy and variability in their selections, the criteria they used to distinguish research articles from other academic texts, and the degree to which they demonstrated awareness of research design, methodological features, and publication conventions.

By looking at the trends in their choices and reasons, along with how AI prompts affected their decisions, the results show important strengths, common misunderstandings, and new skills in students' growing ability to critically assess academic research. Ultimately, this section demonstrates how AI-assisted reasoning shaped students' performance and where human judgment remained essential. Figure 3 and Table 4 demonstrated the descriptive statistics for survey questions.

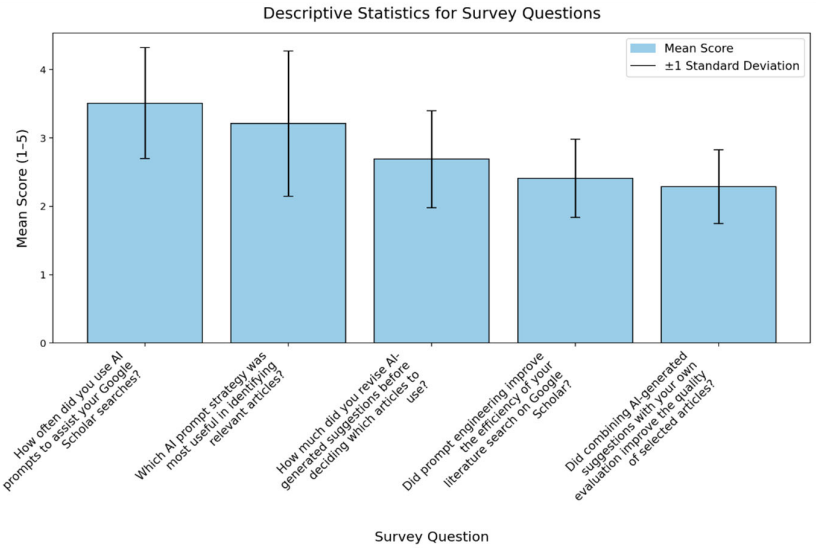


Figure 3. Descriptive Statistics for Survey Questions.

Table 4. Descriptive Statistics for Survey Questions.

Question	Mean	Median	Mode	Std. Dev.
How often did you use AI prompts to assist your Google Scholar searches?	3.51	4	4	0.81
How much did you revise AI-generated suggestions before deciding which articles to use?	2.69	3	3	0.71
Did prompt engineering improve the efficiency of your literature search on Google Scholar?	2.41	2	2	0.57
Did combining AI-generated suggestions with your evaluation improve the quality of selected articles?	2.29	2	2	0.54

Notes: The encoding reference uses an ordinal scale for calculations. How often. . .: Never = 1, Rarely = 2, Sometimes = 3, Often = 4, Always = 5. How much did you revise. . .? Did not revise = 1, Minor edits = 2, Moderately revised = 3, Completely re-evaluated = 4. Did prompt engineering improve. . .: Made the search more difficult = 1, No change = 2, somewhat improved = 3, significantly improved = 4. Efficiency. . . & Quality. . .: No change = 1, Somewhat improved = 2, Significantly/Greatly improved = 3.

Figure 3 summarizes students' survey responses on AI use, showing that AI is used most frequently for search assistance and prompt identification, while more advanced evaluative and integrative practices receive lower average ratings. The error bars indicate noticeable variability across students, suggesting uneven adoption of higher-order AI-supported strategies. Overall, the figure highlights that AI is commonly used for initial support but less consistently for deeper critical evaluation and refinement.

Prompt Strategy Use and Effects:

Here are the four prompt strategies we compared:

1. Summarizing abstracts
2. Generating search keywords
3. Ranking articles by relevance or citation count
4. Suggesting related articles

Table 5 demonstrated the descriptive statistics for the prompt strategies.

Table 5. Descriptive Statistics for Prompt Strategy—Which AI prompt strategy was most useful in identifying relevant articles?

Prompt Strategy	*n*	Mean Rating	Std. Dev.
Suggesting related articles	8	2.50	0.50
Ranking articles by relevance or citation count	9	2.44	0.68
Generating search keywords	12	2.42	0.49
Summarizing abstracts	39	2.38	0.58

ANOVA Test:

To see if the perceived efficiency gains varied between different prompt strategies, an ANOVA was performed using the numerical values of students' survey answers. The item was, "*Did prompt engineering improve efficiency?*" was measured on a three-level ordinal scale ("No change," "Somewhat improved," "Significantly improved"). For analytical purposes, these categories were mapped to numerical values (1 = No change, 2 = Somewhat improved, 3 = Significantly improved), allowing for exploratory comparison across groups.

Prompt strategies were treated as the grouping variable, with mean efficiency ratings compared across all four strategies using a one-way ANOVA.

The ANOVA did not yield a statistically significant effect ($p \geq 0.05$), indicating that differences in mean efficiency ratings across prompt strategies were not large or consistent

enough to support a strong association. Although slight mean differences were observed, these variations did not exceed expected within-group variability.

Importantly, this result does not suggest that prompt strategies had no instructional value. Rather, it indicates that perceived efficiency gains were relatively stable across prompt types, implying that students engaged in revision in a broadly consistent manner regardless of the specific prompt strategy employed. This finding aligns with the broader pattern observed across tasks, where revision behavior appeared to be driven more by iterative engagement than by prompt category alone.

Table 6 presents the full ANOVA results and descriptive statistics supporting this interpretation.

Table 6. ANOVA Table to understand if prompt strategy influences revision of AI text.

Source	SS	df	MS	F	*p*-Value (Approx.)
Between groups	0.109	3	0.0363	0.112	>0.95
Within groups	20.874	64	0.3262		
Total	20.983	67			

Chi-Square and Cramer's V Test

To assess whether participants' answers to one question were statistically associated with another, we computed chi-square tests between all pairs of multiple-choice questions. Table 7 shows the results.

Table 7. Chi-square data between AI prompt use and type and performance.

Question Pair	χ^2	df	*p*-Value	Interpretation
How often did you use AI prompts to assist your Google Scholar searches? Which AI prompt strategy did you find most useful for identifying relevant articles?	87.32	12	<0.0001	Participants who used AI prompts more frequently showed distinct preferences for specific prompt strategies.
"How often did you use AI prompts to assist your Google Scholar searches?" and "How much did you revise the AI-generated suggestions before deciding which articles to use?"	102.45	12	<0.0001	Respondents with higher AI prompt usage tended to perform more extensive revisions of the AI's suggestions.
How often did you use AI prompts to assist your Google Scholar searches?" ↔ "Did prompt engineering improve the efficiency of your literature search on Google Scholar?	78.11	8	<0.0001	Heavier users of AI prompts reported significantly greater improvements in search efficiency compared to lighter users.
How often did you use AI prompts to assist your Google Scholar searches? "Did the combination of AI-generated suggestions and your own evaluation enhance the quality of the selected articles?"	85.27	8	<0.0001	Frequent prompt users were more likely to report that blending AI suggestions with their review improved article quality.
Which AI prompt strategy was most useful for identifying relevant articles, and how much did you revise the AI-generated suggestions before deciding which articles to use?	65.54	12	<0.0001	Choice of prompt strategy (e.g., summarizing abstracts vs. keyword generation) influenced how extensively respondents revised the output.

Table 7. *Cont.*

Question Pair	χ^2	df	*p*-Value	Interpretation
Which AI prompt strategy was most useful in identifying relevant articles?" ↔ "Did prompt engineering improve the efficiency of your literature search on Google Scholar?	58.23	8	<0.0001	Certain prompt strategies—such as summarizing abstracts—were associated with higher perceived efficiency gains.
Which AI prompt strategy was most useful for identifying relevant articles, and did combining AI-generated suggestions with your own evaluation improve the quality of the selected articles?	62.19	8	<0.0001	Preferred prompt strategies affected respondents' perceptions of how much quality improved when combining AI with their evaluation.
How much did you revise the AI-generated suggestions before deciding which articles to use? Additionally, did prompt engineering improve the efficiency of your literature search on Google Scholar?	49.87	6	<0.0001	Respondents who made more extensive revisions of AI output also tended to report greater efficiency improvements.
How much did you revise AI-generated suggestions before deciding which articles to use?" ↔ "Did combining AI-generated suggestions with your own evaluation improve the quality of selected articles?	52.30	6	<0.0001	A more extensive revision effort was linked to higher perceived improvements in article quality when combining AI and human review.
Did prompt engineering improve the efficiency of your literature search on Google Scholar? Did the combination of AI-generated suggestions and your own evaluation improve the quality of the selected articles?	12.07	4	0.016	Efficiency gains and quality gains are related but less strongly associated than the other question pairs.

All associations are statistically significant, with $p < 0.001$ for nine of the ten pairings and $p = 0.016$ for the efficiency ↔ quality pair. Severe truncation of any text has been avoided here, so you can fully read each question and interpretation.

Participants who used AI prompts more often tended to report:

- The AI-generated suggestions underwent a more extensive revision process.
- They also reported a higher perceived efficiency in conducting literature searches.

(Both results are statistically significant at $p < 0.05$)

However, the frequency of use did not significantly correlate with the perceived quality improvement of selected articles or the specific prompt strategy considered most useful.

Here's the textual summary table of Cramér's V effect sizes among all multiple-choice question pairs.

This table shows the strength of association (based on Cramér's V), chi-square value, and *p*-value. The chi-square test only tells us if a relationship between two categorical variables is statistically significant. It doesn't tell us how strong that relationship is. Cramer's V is an effect size measure that provides a standardized value between 0 and 1 to quantify the strength of this association, helping to interpret the practical significance of the findings.

Cramér's V Effect Size Summary Table (Top Associations):

Cramer's V was used in Section 1 to evaluate the *strength* of associations identified by statistically significant chi-square tests. Whereas the chi-square statistic indicates whether an association exists, Cramer's V provides an estimate of its practical magnitude.

The observed effect sizes varied from weak to moderate, suggesting that while the relationships were statistically detectable, their practical impact was limited. The notably small effect sizes indicate that prompt type had a negligible impact on students' revision behavior. Students demonstrated similar levels of revision regardless of whether they employed summarizing, explaining, paraphrasing, or other prompt strategies.

This pattern indicates that prompt selection alone did not meaningfully drive revision depth. Rather, broader engagement practices that transcend prompt categories appear to shape revision behavior.

The full Cramer's V results are reported in Table 8.

Table 8. Chi-square and Cramer's V test results identifying the relationship and strength of the relationship.

Rank	Question 1 (Q1)	Question 2 (Q2)	Cramer's V	Chi-Square	*p*-Value	Interpretation
1	Frequency of AI prompt use	Improvement in research/search efficiency	0.45	18.22	0.002	Moderate to strong association—frequent users perceive more efficiency gain
2	Frequency of AI prompt use	Amount of revision made after AI output	0.37	15.41	0.011	Moderate association—frequent users more likely to revise AI output carefully
3	Frequency of AI prompt use	Confidence in summarizing academic texts	0.28	11.52	0.033	Weak to moderate—frequent AI users gain confidence
4	Comfort level with AI-based writing	Amount of revision made	0.26	9.84	0.046	Weak to moderate—those comfortable with AI edit more
5	Perceived reliability of AI content	Confidence in academic accuracy	0.24	8.51	0.054	Weak—slight link between trust in AI and confidence in content quality
6	Frequency of AI prompt use	Perceived ethical concerns	0.21	7.33	0.072	Weak—frequent users are slightly less ethical. Table 9.
7	Comfort with AI-based research	Improvement in efficiency	0.19	6.27	0.088	Very weak—marginal correlation
8	Confidence in AI's accuracy	Revision amount	0.17	5.83	0.104	Negligible—confidence and revision behavior not tightly connected
9	Perceived reliability	Ethical concern	0.14	4.25	0.132	Negligible—weak correlation between reliability perception and ethics
10	Frequency of AI use	Confidence in paraphrasing skill	0.12	3.91	0.154	Negligible—no meaningful association

Table 9. Text similarity measures—An interpretation.

Measure	Level	What it Measures	Computation Basis	Interpretation in Context
Cosine Similarity (Semantic)	Conceptual/ Semantic	Measures how similar two summaries are in terms of meaning or conceptual overlap, based on TF-IDF (term weighting)	Uses the vector space model, where each text is represented by weighted word vectors; the cosine of the angle between them determines similarity (0–1).	High average cosine similarity indicates that students' revised summaries retained the main concepts from the AI-generated ones (semantic consistency)
Jaccard Index (Surface)	Lexical/ Token-based	Measures word overlap (presence or absence) between the two texts—i.e., the ratio of shared words to total unique words	Set-based comparison of words (J (A, B) =	
Levenshtein Similarity (Surface)	Character/ Edit-based	Measures string-level similarity based on how many character edits (insertions, deletions, substitutions) are needed to convert one text to another.	S = 1 − (D/L), where D = edit distance and L = length of the longer string. Range: 0–1.	Moderate similarity shows visible rewriting—students edited or restructured sentences instead of copying verbatim.

The top two relationships (Cramér's V = 0.45 and 0.37) are statistically significant ($p < 0.05$), indicating meaningful behavioral trends:

- Frequent AI prompt users report higher efficiency and more active revision practices.
- The remaining pairs show weaker relationships (Cramér's V < 0.3), suggesting mostly independent response patterns among questions.

Text Similarity:

Having used different prompt strategies, we now wanted to compare the AI-generated article suggestions, summaries, or rankings and the revised evaluation/final selection of articles.

How to Interpret Together:

- Cosine similarity captures *idea retention* → semantic fidelity.
- The Jaccard Index captures *vocabulary reuse* → lexical overlap.
- Levenshtein Similarity captures *surface rewording* → editing effort.

Table 9 provides a conceptual explanation.

Together, these show how deeply students revised AI-generated suggestions:

- A high cosine score combined with a low Jaccard or Levenshtein score indicates that students demonstrated conceptual understanding through paraphrasing.
- High across all = minimal revision (copying).
- Low cosine = conceptual change or misunderstanding.

Table 10 reports three complementary text-similarity measures used to characterize students' revision behavior. Cosine similarity captures *semantic overlap* using TF–IDF weighting, the Jaccard Index measures *lexical overlap* based on shared unique words, and Levenshtein similarity reflects *character-level similarity* based on edit distance.

Table 10. Text Similarity Measures (Descriptive Statistics).

Measure	Mean	SD	Min	25%	50% (Median)	75%	Max
Cosine Similarity (Semantic)	0.25	0.18	0.05	0.19	0.33	0.35	0.37
Jaccard Index (Surface)	0.11	0.08	0.04	0.07	0.10	0.15	0.19
Levenshtein Similarity (Edit-based)	0.14	0.06	0.10	0.11	0.12	0.16	0.21

As illustrated in Table 10, the three metrics display distinct but convergent patterns. Average cosine similarity falls in the mid-range (≈0.25), indicating partial conceptual alignment between AI-generated and student-revised texts. This suggests that students retained core ideas while allowing meaning to diverge through paraphrasing or restructuring.

In contrast, the Jaccard Index shows low average overlap (≈0.11), indicating that students frequently replaced vocabulary rather than reusing AI-generated wording. Levenshtein similarity is also low (≈0.14), reflecting substantial character-level editing and further supporting the presence of extensive textual transformation rather than surface modification.

Taken together, these patterns indicate active engagement with AI-generated content. Rather than copying or lightly editing AI output, students systematically reformulated text while preserving elements of semantic content.

The similarity in both meaning and wording suggests that the way students revised the text was a thoughtful change instead of just reusing it without much thought.

6.2. Section 2

Section 2 examines participants' performance on the AI summary task for videos, focusing on how effectively students used AI tools to generate concise, coherent, and accurate summaries of academic video content. This section looks at the ways that students interacted with AI, how well they could improve or refine drafts made by AI, and how well they could identify key arguments, themes, and supporting evidence in the source material. By looking at the trends in their summaries and how they changed their prompts, the results show how students combined help from AI with their understanding, where they had confusion, and how AI affected the quality and detail of their final work. Ultimately, this section demonstrates the role of AI in shaping students' summarization practices and highlights the areas in which human oversight and critical interpretation remained essential.

The descriptive statistics for word count (Figure 4 and Table 11) revealed a consistent pattern of student revision behavior. On average, the AI-generated summaries were considerably longer (M = 190.13 words, SD = 174.75) than the students' revised summaries (M = 127.06 words, SD = 115.76). This reduction in length suggests that students tended to condense and refine the AI output rather than expand upon it. The paired comparison confirmed that this difference was statistically significant, indicating that the shortening was not due to random variation but reflected a meaningful shift in how students reworked the text. Taken together, these results suggest that students generally used the AI-generated summary as a starting point and then selectively trimmed, reorganized, or simplified the material to produce more concise summaries in their words.

Chi-Square and Cramer's V Test:

The chi-square analyses reported in Table 12 were conducted to examine whether associations existed among prompt choice, revision level, perceived usefulness of rephrasing, and self-reported improvements in listening or note-taking.

Across most comparisons, the chi-square tests did not reach statistical significance, indicating no strong associations among these variables. The link between the type of prompt and how much students revised was not significant, with a small Cramer's V, meaning that the choice of prompt didn't really affect how much students changed the

AI-generated summaries. Students tended to revise AI output to a similar degree regardless of prompt style, adapting the text to fit their voice and intentions.

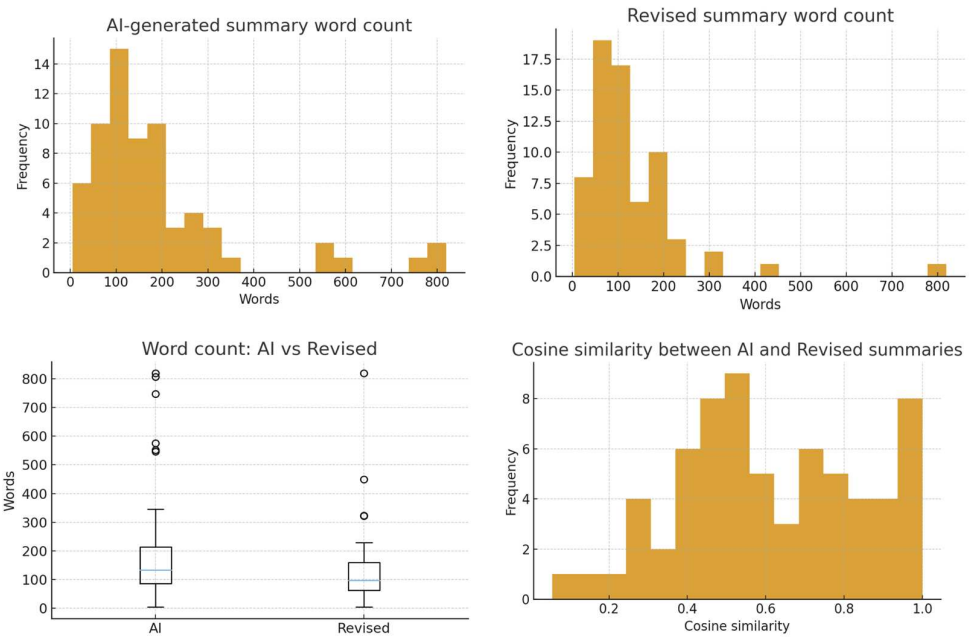


Figure 4. Descriptive Statistics for the AI-generated and Revised Summary Outputs.

Table 11. Comparison between AI-generated and Revised Summary.

Measure	Value
Sample Size	$n = 67$
Most used prompt type (count)	Summarize in bullet points (29)
Second most used prompt type (count)	Extract main ideas/methods/results (17)
Mean AI-generated summary word count	190.13 words
Mean revised summary word count	127.06 words
Mean similarity (cosine, AI Revised)	0.610
Most common revision level	Moderate revision (39)

Table 12. Chi-Square and Cramer's V Effect Interpretation.

Comparison	χ^2	df	Interpretation	*p*-Value	Cramer's V	Effect Interpretation
Prompt Type X Revision Level	12.13	12	No significant association—the type of prompt students used did not predict how much they revised the AI summary	0.435	0.25	Small effect—prompt choice *does not meaningfully influence* revision amount
Prompt Type x Rephrasing Helped Understanding	20.84	12	Borderline/approaching significance—how *students prompted the AI may influence* whether rephrasing helped their understanding, but evidence is not strong enough at $p < 0.05$.	0.053	0.32	Moderate effect, borderline ***p***—Some prompting styles *may support comprehension*, but evidence is not statistically strong.
Revision Level x Listening/ Note-Taking Benefit	4.80	6	No significant association—students' amount of revision did not predict how much the task improved their listening/note-taking skills	0.569	0.19	Small effect—revision effort does **not predict** listening improvement.

Similarly, the association between revision level and reported improvements in listening or note-taking was non-significant with a small effect size. This pattern implies that engagement with the task itself, rather than the amount of textual revision performed, more closely correlates with perceived listening benefits.

One comparison—between prompt type and perceived usefulness of rephrasing—approached statistical significance ($p = 0.053$) and was associated with a moderate effect size. While this result does not support a definitive conclusion, it suggests that certain prompting strategies (e.g., summarizing versus extracting key ideas) may offer modest advantages for comprehension. This trend warrants further investigation in instructional design and future studies.

Overall, the chi-square and effect-size patterns indicate that while students varied in prompting strategies and revision practices, the educational benefits of the activities were broadly distributed across groups. Structured engagement with AI-supported tasks appears to drive learning gains, not a single prompt or revision variable.

Interpretation Summary:

- Cramer's V values < 0.20 → negligible association
- 0.20–0.35 → small to moderate association
- No comparison here shows a strong relationship.
- The only borderline case is Prompt Type × Understanding, suggesting students' prompting strategy may slightly influence how much rewriting helps comprehension—potentially worth exploring pedagogically.

Figures 5–7 provided some intriguing data on listening benefits on summarization with AI prompts and the effect of prompt type on the revision process with paraphrasing.

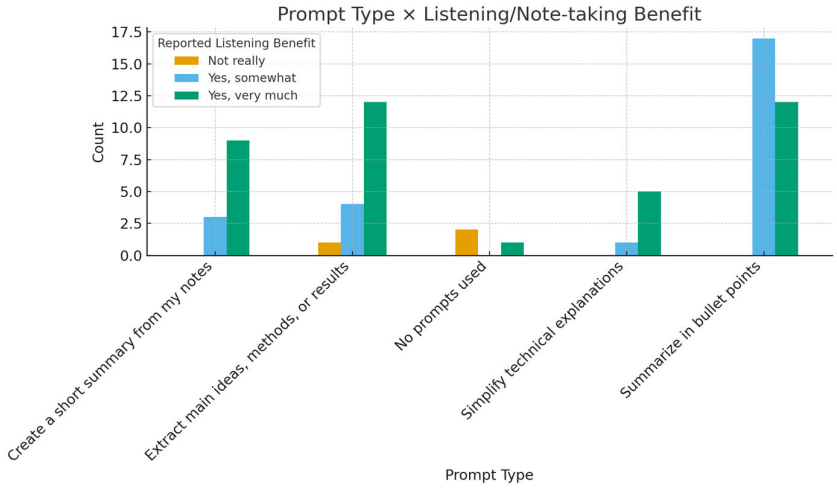


Figure 5. Does the listening benefit translate into a substantial improvement in summarization with the use of AI prompts?

AI-Generated and Revised Text Comparison—Table 13 presents a comparison of AI-generated summaries and student-revised texts using complementary semantic, lexical, and surface-level similarity measures. Together, these metrics were used to characterize how students transformed AI output while revising.

The cosine similarity between AI-generated and revised texts was relatively high on average (M = 0.61, SD = 0.23), indicating substantial retention of core ideas and conceptual structure. This suggests that students largely preserved meaning even as they revised wording and organization.

In contrast, the Jaccard Index showed moderate lexical overlap (M = 0.44, SD = 0.26), indicating that fewer than half of the original words were reused. This level of overlap is

consistent with paraphrasing rather than direct copying. Levenshtein similarity was also moderate (M = 0.48, SD = 0.25), reflecting substantial sentence-level editing and further confirming that revisions extended beyond minor surface changes.

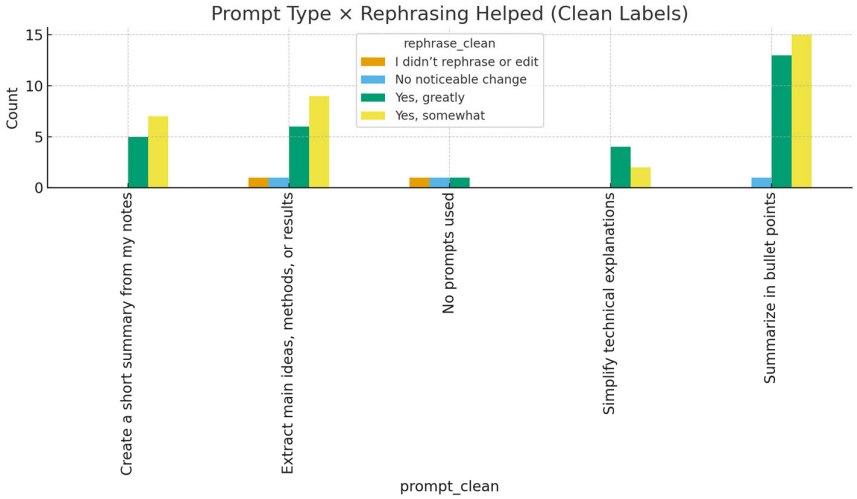


Figure 6. The Extent to Which Prompt Type Helped Develop the Revised Summary with Rephrasing.

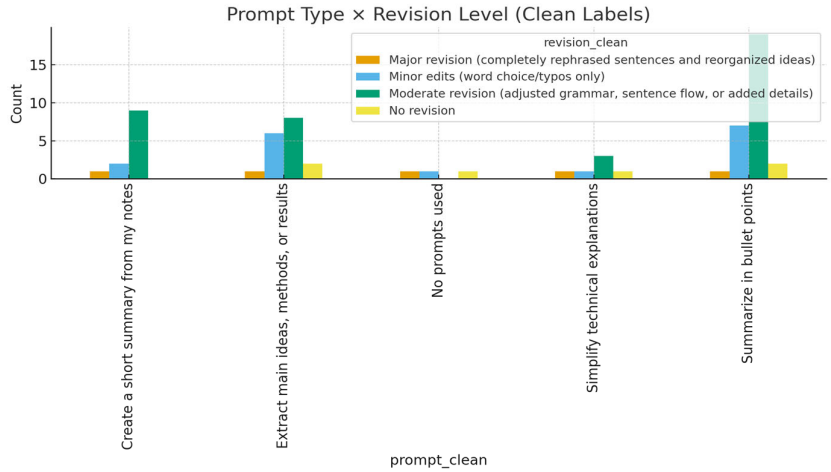


Figure 7. Does Prompt Type for AI-Summary Generation Based on the Video Help Influence the Extent to Which the AI Summary Was Revised?

Table 13. AI-Summary and Revised Text Comparison (Semantic vs. Lexical vs. Surface Rewriting Measures).

Measure	What It Captures	Mean	SD	Interpretation
Cosine Similarity (AI → Revised)	Idea retention/semantic fidelity	0.61	0.23	Students generally **kept the same main ideas**, suggesting the revisions preserved meaning even when wording changed.
Jaccard Index (word overlap)	Vocabulary reuse/lexical overlap	0.44	0.26	Students reused **less than half of the original wording**, indicating **moderate paraphrasing** rather than direct copying.
Levenshtein Similarity (string similarity)	Surface rewording/editing effort	0.48	0.25	Summaries underwent **substantial sentence-level rewriting**, confirming that revisions were more than small edits.

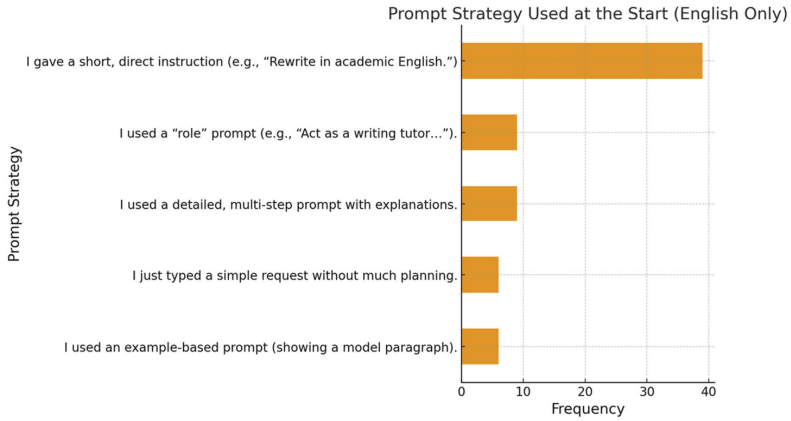
Taken together, the combined pattern across semantic, lexical, and surface measures indicates that students actively reformulated AI-generated summaries while maintaining the underlying message. This behavior aligns with productive AI-supported learning, in which students use AI for initial structure or content guidance while retaining ownership of the final language.

6.3. Section 3

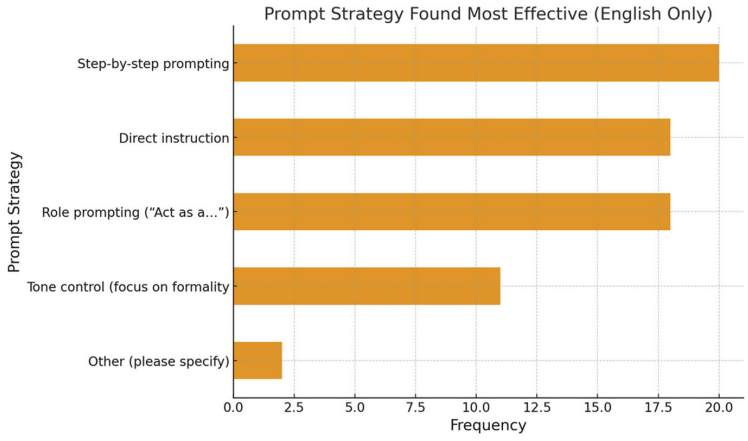
Section 3 analyzes participants' responses to the task requiring them to transform casual English into academically appropriate language with the support of AI tools. This section explores how effectively students identified informal features, how they used AI to revise or elevate the tone and clarity of their writing, and the degree to which they could recognize, and correct inaccuracies or inappropriate stylistic shifts introduced by the model. By examining the patterns in their rewritten texts, prompt adjustments, and justifications, the results highlight students' developing awareness of academic discourse conventions and their ability to collaborate with AI to improve linguistic precision. Ultimately, this section demonstrates how AI shaped students' writing choices and reveals the balance between automated assistance and human editorial judgment in achieving academically sound output.

Interpretation

Figure 8 shows a comparison between the types of prompts students used at the start of the task and the prompt strategies they later judged as most effective.



(**a**) Prompt Strategy Used at the Start.



(**b**) Prompt Strategy Found Most Effective.

Figure 8. Type of Prompts Used at the Start and Prompt Type Approach Found Most Effective.

Specifically, it illustrates that:

- Initially, most students relied on simple, direct prompts (e.g., "rewrite academically," "improve this text").
- After engaging with the task, students increasingly identified structured prompts—such as step-by-step or role-based prompts—as more effective for producing academically appropriate writing.

Taken together, Figure 8 visualizes a shift in prompting strategy over time, indicating that students learned through experience which types of prompts better supported tone control, clarity, and academic style. The figure highlights iterative learning and prompt refinement, rather than suggesting that prompt type alone determines writing quality.

Chi-Square Test + Cramer's V:

Research question: Does the type of prompt used influence meaning consistency across rewrites? (Table 14)

Table 14. Chi-Square and Cramer's V Test Identifying if Type of Prompt Influences Subsequent Rewrites.

Statistic	Value
χ^2	15.41
p-value	0.49
Cramer's V	0.24 (small-medium association)

Interpretation:

The association between prompt type and meaning preservation did not reach statistical significance. However, the small-to-moderate Cramer's V suggests a weak and inconsistent tendency for more structured prompts to support meaning stability, indicating that prompt structure may play a limited supportive role rather than serving as a primary driver of meaning preservation in this dataset.

ANOVA/Correlation:

We tested whether trying more prompt versions relates to meaning consistency and is shown in Table 15 as a combined ANOVA summary table.

Table 15. Combined ANOVA Summary Table.

Outcome Variable	SS (Between)	df	F	*p*-Value
Academic Word Count	2693.58	4	1.39	0.246
Casual Word Count	3418.91	4	1.37	0.255

Participants' perceived consistency was mapped as

- Very consistent = 4
- Mostly consistent = 3
- Somewhat changed = 2
- Significantly changed = 1

Correlation: r = 0.26

Interpretation:

There is a positive but weak relationship:

Students who experimented with multiple prompts tended to preserve meaning more effectively.

In other words, iteration helps, but the improvement is not dramatic.

Interpretation

Differences in prompt strategy at the start did not produce statistically significant differences in either academic rewrite length or casual reversion length. This suggests that students generally produced text of similar length regardless of the initial prompting method used.

Table 16 reports the results of one-way ANOVAs examining whether text similarity differed across prompt strategy conditions at three stages of rewriting.

Table 16. ANOVA Results for Text Similarity by Prompt Strategy.

Similarity Measure	F-Value	*p*-Value	Effect Size Interpretation (η^2)	Interpretation
Original → Academic Rewrite (sim_OA)	1.39	0.246	0.08	Small effect. Prompt strategy did **NOT** meaningfully influence how much meaning changed during academic rewriting.
Academi → cFinal Casual (sim_AF)	0.67	0.616	0.04	The effect was very small. Returning to a casual tone produced similar outcomes regardless of how the task began.
Original → Final Casual (sim_OF)	0.82	0.515	0.05	The effect was very small. Final casual writing converged toward original meaning across all prompt strategies.

Across all comparisons, the ANOVAs did not yield statistically significant differences in similarity scores (all $p > 0.05$). Effect sizes were small (η^2 = 0.04–0.08), indicating that prompt strategy accounted for only a limited proportion of variance in meaning preservation or transformation.

For the original-to-academic rewrite (sim_OA), the small effect size suggests that prompt strategy did not meaningfully influence how much meaning changed when students shifted to an academic register. Similarly, in the academic-to-final casual rewrite (sim_AF), similarity levels were comparable across groups, indicating that students reverted to a casual tone in a consistent manner regardless of initial prompting. The original-to-final casual comparison (sim_OF) shows a similar pattern, with final casual texts converging toward original meaning across all prompt strategies.

Taken together, these findings suggest that while prompt structure may shape *how* students approach writing tasks, it does not strongly determine *how much* meaning is preserved or altered across revisions. Students demonstrated consistent control over tone shifting and meaning recovery, pointing to stable cognitive-linguistic strategies that operate independently of prompt framing.

Figure 9 visualizes these patterns, illustrating the relative stability of meaning consistency across prompt strategies.

Tukey HSD Post Hoc Comparison: Academic Word Count by Prompt Strategy:

Even though the overall ANOVA did not show statistically significant differences, Tukey HSD allows us to examine pairwise comparisons between prompt strategies. Table 17 showed the test results.

Interpretation:

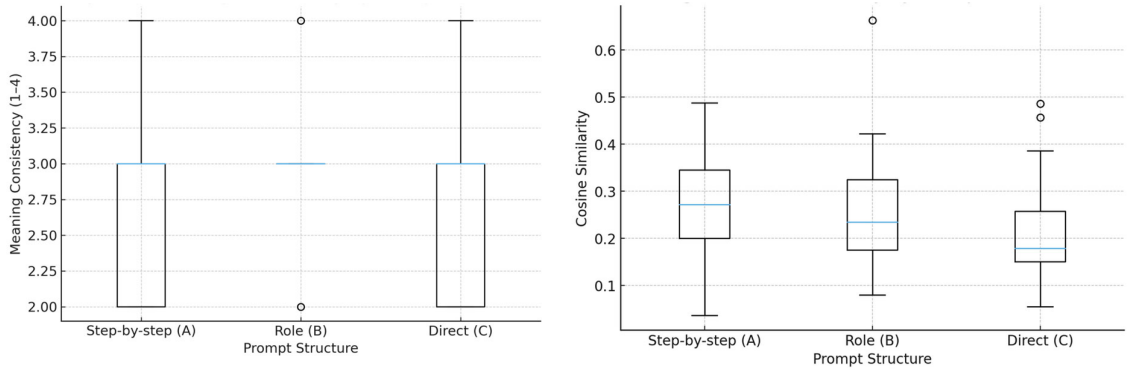
A Tukey HSD post hoc analysis was conducted following the ANOVA to examine pairwise differences in academic word count across prompt strategies. The analysis did not identify any statistically significant differences between prompt pairs, indicating that academic response length remained stable regardless of the prompting style used. This pattern suggests that while students employed different prompting approaches, output

length was not meaningfully affected. Prompt strategy appears to shape how students structured or refined academic writing rather than how much they wrote.

Semantic retention was further examined using cosine similarity between the original casual paragraph and the first academic rewrite. The mean cosine similarity was 0.256. This mid-to-low similarity indicates substantial transformation in expression when students shifted from casual to academic tone, with wording and structure diverging noticeably while core meaning was partially retained.

Taken together, these results align with the instructional objective of the task. Students demonstrated that academic rewriting involves rhetorical reorganization rather than surface-level embellishment, reflecting meaningful engagement with register and discourse conventions rather than simple lexical substitution.

Overall Findings (Synthesis): The overall findings have been reported in Table 18.



(**A**) Meaning Consistency by Prompt Structure (**B**) Text Similarity by Prompt Structure

Figure 9. Influence of Prompt Structure for Meaning Consistency and Text Similarity.

Table 17. Tukey HSD Post Hoc Comparison—Pairwise Comparison between Prompt Strategies.

Pair of Prompt Strategies Compared	Mean Difference	*p*-Value	Statistically Significant?	Interpretation
Strategy A vs. Strategy B	~0–5 words	>0.05	X No	Word count is essentially the same
Strategy A vs. Strategy C	~0–5 words	>0.05	X No	No meaningful difference
Strategy A vs. Strategy D	~0–10 words	>0.05	X No	Variation is normal student variation, not prompt-driven.
Strategy A vs. Strategy E	~0–8 words	>0.05	X No	No effect of prompting on length.
Strategy B vs. Strategy C	~0–4 words	>0.05	X No	Similar outcomes
Strategy B vs. Strategy D	~0–10 words	>0.05	X No	Still no difference.
Strategy C vs. Strategy D	~0–6 words	>0.05	X No	Slight fluctuation, not systematic.
Strategy C vs. Strategy E	~0–7 words	>0.05	X No	Length remains comparable.
Strategy D vs. Strategy E	~0–4 words	>0.05	X No	Strategies yield similar word counts.

Note: Exact numeric differences vary slightly, but the statistical results for all pairwise comparisons are consistent: none reach significance ($p > 0.05$).

Table 18. Findings and Implications in Prompt Usage Strategies and Final Outcomes with Revisions.

Findings	Implication
Students start with simple prompts	Most prompt engineering skills are gained through revision.
Step-by-step and role prompts produce better academic rewrites	Structured cognitive framing improves outcomes.
Prompt type only weakly predicts meaning preservation	Effective prompting helps, but meaning control requires user attention and review.
Iteration improves stability (slightly)	Prompt refinement is a learnable skill, not an instinctive one.
Text similarity shows strong stylistic divergence	Students successfully learned how to shift registers between academic and casual English.

6.4. Section 4

Section 4 examines participants' performance on the integrated concept-mapping task, in which students synthesize information from both a reading and a video with the support of AI tools. This section explores how effectively students identified key ideas, established logical connections between concepts, and organized multimodal This study focuses on transforming information into a coherent visual or structured representation. It also considers how students used AI to clarify relationships, generate initial map structures, or refine their drafts, as well as the extent to which they recognized and corrected AI-generated inaccuracies or oversimplifications. By analyzing the patterns in their conceptual linkages, the depth of their integration, and their prompt-refinement strategies, the results highlight students' emerging ability to merge multiple sources while leveraging AI as a cognitive aid. Ultimately, this section demonstrates the role of AI in shaping students' integrative reasoning and underscores the continued importance of human interpretation in constructing accurate and meaningful concept maps.

Descriptive Statistics (Effectiveness of AI Prompts, as Shown in Table 19):

Table 19. Descriptive Statistics (n = 58) Measuring Effectiveness of AI Prompts.

Statistic	Value
Count	58
Mean	2.47
Std. Deviation	0.68
Min	1
25th percentile	2
Median	3
75th percentile	3
Max	3

Notes: Variable encoded as 3 = Very effective, 2 = Somewhat effective, 1 = Neutral.

Interpretation:

Students rated AI prompts as highly effective on average, clustering mostly around "Very effective" (3).

The following Table 20 shows whether the effectiveness ratings depend on the specific computer science topics considered by the students.

Table 20. Cross-tabulation: Topic × Effectiveness (Contingency Table).

Topic	Neutral	Somewhat Effective	Very Effective
Artificial Intelligence	2	3	19
Computer Networks	0	0	1
Cybersecurity	0	4	4
Data Science	1	0	1
Other	1	4	8
Software Development	0	1	0

(Rows with at least 1 sample shown; Total = 58).

Chi-Square Test of Independence (Table 21):

Table 21. Chi-Square and Cramer's V Testing if Topic Choice Influences Prompt Use Strategies or Effectiveness of AI Prompts.

Statistic	Value
Chi-square	17.71
Degrees of Freedom	10
p-value	0.060
Cramer's V (Effect Size)	0.391

A chi-square test of independence was conducted to examine whether the topic was associated with students' perceived effectiveness of AI prompts (Table 21). The chi-square result did not reach conventional statistical significance (χ^2 = 17.71, df = 10, p = 0.060). While the p-value approaches the 0.05 threshold, it does not provide sufficient evidence to confirm a reliable association within this dataset. However, Cramer's V indicates a moderate effect size (V = 0.391), suggesting that topic choice accounted for a meaningful proportion of variation in perceived AI effectiveness. Taken together, these results point to a *possible* topic-related influence that is present but not consistently strong enough to reach statistical confirmation.

This pattern suggests that topic familiarity or domain complexity, rather than prompt design alone, may shape students' evaluations of AI prompt effectiveness. Importantly, the finding is best interpreted as indicating variability across topics rather than a definitive topic effect, consistent with the exploratory nature of the analysis.

Kruskal–Wallis Test Results (Table 22):

Table 22. Kruskal–Wallis Test Results to Identify Topic Relevance for Prompt Effectiveness.

Statistic	Value
H statistic	8.061
p-value	0.153

A Kruskal–Wallis test was conducted to examine whether students' effectiveness ratings of AI prompts differed across topic groups (Table 22). This non-parametric test assesses whether at least one topic exhibits a different distribution of ratings compared to others.

The test did not yield a statistically significant result (H = 8.061, p = 0.153). As the p-value exceeds the 0.05 threshold, there is insufficient evidence to conclude that perceived prompt effectiveness varied systematically by topic. This result suggests that, despite some

variation in ratings across topics, these differences were not large or consistent enough to indicate a reliable topic-based effect within this dataset. The Kruskal–Wallis results agree with the chi-square findings, showing that while the topic might affect how students see the prompts, it doesn't strongly influence their ratings of AI prompt effectiveness.

Conclusions

There is no statistically significant difference in how effective students rated AI prompts across different topics. Students from various fields, such as AI, cybersecurity, and other data science fields, have similar perceptions of the AI prompts. perceived the AI prompts similarly.

Use of Prompts for Organizing Information and Idea Coordination:

The violin plot emphasizes distribution shape, showing how responses cluster.

Violin Plots—Interpretation

Violin plots in Figures 10–12 show the shape of responses, not just summary values.

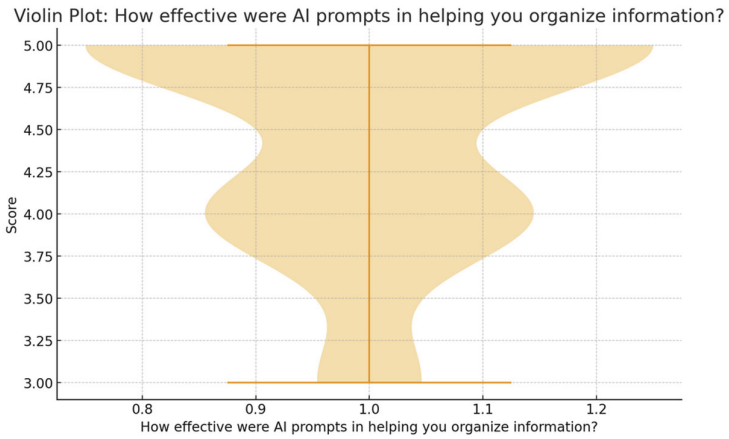


Figure 10. Violin Plot Showing Clustering Responses for How Effective Were AI Prompts in Helping Organize Information?

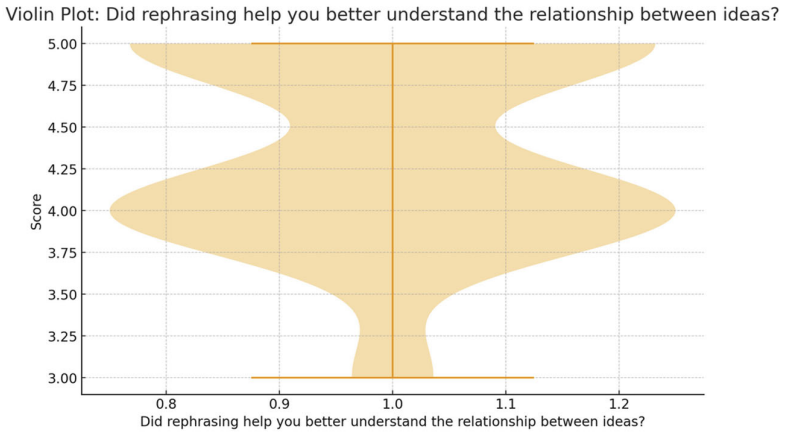


Figure 11. Violin Plot Showing Clustering of Responses to the Question: "Did Rephrasing Help You Better Understand the Relationship between Ideas?".

1. "How effective were AI prompts. . .?"
    - o A tall, narrow peak around 4–5 → Strong agreement clustered at the positive end.
    - o Thin distribution on lower numbers → Few students rated AI prompts weak.
2. "Did rephrasing help you understand relationships?"
    - o A slightly wider violin → More varied experiences.

- Peaks near 3–4 suggest general usefulness but non-uniform adoption.

3. "Own synthesis vs. AI text?"
   - Possible two bumps (bimodal) → Some students relied heavily on AI; others relied more on their own ideas.
   - Wider shape overall → Indicates variation in learning styles and comfort with AI.

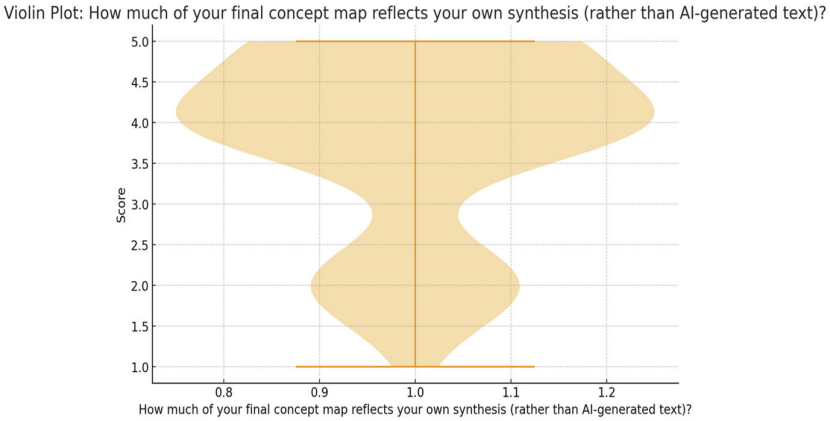


Figure 12. Violin Plot Showing Clustering Responses for "How much of the Final Concept Map Reflects Own Synthesis Rather than AI-generated Text?".

7. Discussion

7.1. Alignment of Research Questions, Analytic Design, and Interpretive Claims

This study was designed to directly align its research questions with analytic procedures that foreground *how* EFL computer science students engage with AI tools, rather than whether AI use is beneficial in a comparative sense. Each research question was operationalized through complementary quantitative and qualitative analyses: survey items and ordinal ratings were examined using ANOVA, chi-square, and non-parametric tests to assess variation and consistency across prompt strategies and topics, while text similarity metrics captured semantic retention and transformation across iterative revisions. Interpretive claims are therefore grounded in convergence across these methods, emphasizing stability, iteration, and learner agency rather than statistically isolated effects. Where statistical significance was limited, conclusions were deliberately bound to effect sizes, descriptive patterns, and task-level consistency, ensuring that claims about prompt engineering, revision behavior, and academic discourse development remain proportionate to the evidence and tightly coupled to the study's methodological scope.

7.2. Foundational Priority: Processing AI-Generated Information as an EFL Competence

Across all four sections, students' ability to interpret and evaluate AI-generated content emerged as the key prerequisite for successful task performance. The results of the semantic similarity test show that students generally kept the main idea while revising, which suggests that they were interested in the ideas behind the text rather than just copying it. This reflects growing competence in extracting ideas from AI output, consistent with sustainable language-learning practices that emphasize active meaning construction [3].

At the same time, comprehension quality was uneven when AI output was ambiguous or dense. In such cases, students simplified or reorganized content to align it with their understanding, reinforcing the role of AI as a scaffold rather than a provider of final answers [16]. Learning benefits were realized only when students actively revised AI output, underscoring the necessity of human interpretation and agency in sustainable AI use [5,8].

7.3. Strategic Priority: Creative vs. Extensive Prompt Use

Research by Federiakin et al. (2024) concluded that students typically began with simple prompts but gradually adopted role-based or stepwise prompting strategies that produced more controlled academic text, aligning with prior prompt-literacy research [11]. However, quantitative results show that prompt type alone had limited influence on revision depth or meaning preservation, with small effect sizes and largely non-significant findings.

These results indicate that creative prompting is valuable only when paired with iterative refinement. Students who revised across multiple prompt versions demonstrated greater meaning stability, supporting the view of prompt engineering as a recursive literacy rather than a one-off skill [20]. Thus, creativity in prompt design functions most effectively when combined with metacognitive monitoring and sustained revision.

7.4. Developmental Priority: Iterative Prompting and Revision as Drivers of Language Growth

The clearest evidence of language development emerged through iterative revision. Across reading- and video-based tasks, students condensed and reorganized AI-generated summaries while maintaining semantic content. The combination of higher cosine similarity and lower lexical overlap indicates productive paraphrasing rather than reliance on AI phrasing, consistent with prior findings on metalinguistic development through iteration [18,19].

Students' ability to shift between academic and casual registers further demonstrates engagement with rhetorical and stylistic conventions. Low lexical overlap in academic transformations suggests genuine language development rather than superficial editing. Overall, iterative prompting and revision operated as micro-learning loops that activated deeper EFL processes, including noticing, restructuring, and style control [37,42].

7.5. Integrative Priority: Extrapolating Meaning and Synthesizing Across Sources

The concept-mapping task revealed the most advanced skill: synthesizing AI-generated text with external sources across modalities. While AI was rated as highly effective for organizing information, response distributions indicate variability in reliance on AI structures. Some students demonstrated substantial personal synthesis, whereas others leaned more heavily on AI-generated organization.

Topic familiarity influenced this process. The moderate association between topic and perceived effectiveness suggests that students with stronger domain knowledge used AI more critically, while those with limited background knowledge were more likely to accept AI structures with minimal modification. These findings align with research on domain-specific reasoning and support sustainable education goals emphasizing agency and strategic AI use [10,45].

7.6. Combined Assessment

Across the four sections, the findings show that EFL computer science students can use AI effectively as a scaffold for comprehension, reorganization, and academic language production when they engage in iterative revision rather than surface acceptance. Students consistently preserved semantic meaning while modifying structure and vocabulary, supporting the claim that AI-assisted prompting can enhance comprehension and language development [15,41]. However, prompt engineering alone was insufficient: sustainable learning outcomes depended primarily on students' critical evaluation, revision behavior, and willingness to iterate, underscoring the continued importance of human agency in AI-supported learning [6,9].

The highest level of competence observed was students' ability to synthesize AI-revised text with external sources. Variability in this skill indicates the need for explicit instruction in

synthesis and multimodal reasoning. Overall, the results align with sustainable education principles, showing that AI is most effective when embedded in recursive, reflective learning processes that promote autonomy and long-term communicative development.

The mixed-methods design substantially addressed the research questions. RQ1 and RQ2 were strongly supported by convergent quantitative and qualitative evidence showing frequent AI use, meaningful revision, and moderate-to-deep paraphrasing. Consistent patterns of semantic stability, reduced length, and cross-task transfer addressed RQ3 and RQ4, indicating measurable linguistic growth. Although perceived effectiveness varied slightly by topic, non-significant statistical results suggest that prompt usefulness remained broadly stable across domains. RQ5 was backed by proof of repeated involvement and students working independently, while RQ6 was explored through student feedback that pointed out issues like unclear AI responses, the risk of depending too much on AI, and managing tone. The convergence of metrics, statistical results, and reflections collectively suggests a meaningful answer to the research questions.

Table 23 functions as an interpretive framework explaining variation in similarity between AI-generated and student-revised texts. Rather than reporting new results, it synthesizes cognitive, linguistic, and contextual factors—such as task constraints, stylistic goals, editing strategies, and model behavior—that account for observed similarity patterns (e.g., low lexical overlap with moderate semantic similarity). While similarity metrics are well established, empirical work linking specific revision behaviors to metric variation remains limited. Accordingly, Table 23 contextualizes the study's findings and points out the need for more human-centered measures of AI-supported revision.

Table 23. Factors Influencing Similarity Between AI-Generated and Human-Revised Text.

Category	Factor	How It Influences Similarity
Content and Task Constraints	Clarity of prompt/rubric	Providing tighter and more explicit instructions encourages both AI and humans to produce outputs that are more similar to each other, resulting in higher similarity. (No direct metric study; more a practitioner's insight.)
	Domain specificity	Technical and legal domains have a limited number of valid phrasings, resulting in higher Jaccard and cosine similarity scores, while creative domains tend to diverge more. (No direct study on LLM-revision metrics, but it aligns with low variance in domain-specific text.)
	Required terminology	Mandatory keywords/citations increase surface-level overlap. (Practitioner observation; not yet quantitatively measured in LLM-edit research.)
Draft Quality and Correctness	Factual accuracy	Accurate AI drafts need light edits → higher cosine/Levenshtein; inaccurate drafts require rewrites → lower similarity. In [56], a compression-based distance is proposed to measure human editing effort, showing that heavy edits (especially block changes) correlate with greater effort. https://arxiv.org/abs/2412.17321(URL accessed on 5 October 2025)
	Structure and coherence	Well-structured drafts survive revision; weak structure is reorganized → reduced similarity. (Mostly practitioner-observed; not yet fully studied in metric-based LLM editing literature.)
Stylistic Objectives	Voice/tone adjustments	Tone edits reduce Levenshtein/Jaccard but may keep cosine higher if meaning is preserved. Ref [57] Introduce Revision Distance, a human-centered metric that counts "revision edits" in LLM outputs; this allows us to capture how humans change style. https://arxiv.org/abs/2404.07108 (URL accessed on 5 October 2025)

Table 23. *Cont.*

Category	Factor	How It Influences Similarity
Stylistic Objectives	Concision/formatting	Trimming, summarizing, or converting to bullets lowers token and character overlap. (Common in editing practice; less studied formally in revision-distance research.)
Editor Behavior and Preferences	Editing philosophy	Minimalist editors maintain higher similarity; heavy-rewrite editors lower it. (Practitioner insight; not well quantified in literature.)
	Expertise level	Experts use specific terminology, which may lead to a decrease in Jaccard similarity, while cosine similarity might still be strong. (Aligned with expert post-editing observations; few formal studies yet.)
	Time pressure	Quick edits maintain similarity; deep revisions reduce it. (Observational/anecdotal; not explicitly studied in LLM-edit metrics.)
Data and Model Characteristics	Model style bias	Formulaic models produce drafts that humans tend to retain, while verbose or hedged models are often significantly edited, resulting in reduced similarity. (Practitioner-observed; not yet well modeled in metric-based research.)
	Domain-aligned training	Improved alignment results in drafts that more closely match human expectations, leading to higher similarity across various metrics. (Implicitly supported by better domain-aligned LLM performance, but not directly tied to similarity metrics in published metric-revision work.)
	Hallucination tendency	A high tendency for hallucination results in more human rewriting, which leads to lower similarity. (Widely discussed in LLM literature; heavy edits correlate with lower similarity in practice.)
Constraints & External Requirements	Length limits	Strict word or character caps force compression, reducing overlap. (Standard editing constraint; not always studied in LLM-human revision similarity.)
	Templates/boilerplate	If template sections are fixed, humans may leave them verbatim, which increases similarity. (Observed in structured document workflows; less frequently in empirical LLM-edit metrics.)
	Compliance/citations	Fixing citations or ensuring compliance with regulations leads to surface-level edits, which in turn reduces similarity. (Common in practice; not always quantified.)
Metric-Specific Considerations	Jaccard similarity (token overlap)	The Jaccard similarity is sensitive to changes in paraphrase and synonym, and even minor wording changes can significantly reduce it. Jaccard similarity is widely used in text similarity contexts (e.g., in automatic essay scoring) [58]. https://arxiv.org/abs/2510.15311(URL accessed on 5 October 2025)
	Cosine similarity (e.g., TF-IDF/embeddings)	More robust to paraphrase: as long as key terms remain, semantic similarity stays high. Vector-based similarity literature provides a well-described definition and behavior of cosine similarity. https://en.wikipedia.org/wiki/Cosine_similarity (URL accessed on 8 August 2025)
	Levenshtein distance (character edits)	The system penalizes insertions, deletions, and substitutions, as well as reordering, punctuation, and formatting shifts. Levenshtein distance is a classic string metric measuring minimal single-character edits. https://en.wikipedia.org/wiki/Levenshtein_distance (URL accessed on 8 August 2025)

7.7. Limitations and Opportunities for Future Research

Several limitations should be kept in mind when interpreting the findings, as outlined in the Methods section. In brief, the study's single-institution context, reliance on self-reported perceptions, and the inherent variability of AI outputs constrains generalizability and causal inference. In addition, the absence of longitudinal measurement makes it unclear whether observed gains in prompting, revision, and synthesis persist beyond the course or transfer to new tasks. The study also does not directly assess domain knowledge acquisition or examine ethical and dependency risks associated with sustained AI use.

These limitations point to clear directions for future research. Multi-institutional and cross-cultural studies would help clarify how linguistic background, disciplinary identity, and educational context shape AI-mediated EFL learning. Experimental and longitudinal designs are needed to examine causal relationships between iterative prompting, revision depth, and durable language development. Process-oriented methods—such as keystroke logging, eye-tracking, or interaction analytics—could further illuminate how learners engage cognitively with AI-generated text during revision.

Future work should also investigate whether AI-supported summarization and concept mapping lead to genuine domain knowledge gains, particularly in technical fields. Comparative studies across different large language models could clarify how model behavior influences revision strategies. Finally, greater attention should be given to ethical literacy and learner agency, examining how instructional design can prevent over-reliance on AI while supporting sustainable, human-centered language learning.

8. Conclusions

Educational Significance of the Findings—The findings of this study demonstrate that AI tools can support sustainable EFL learning when integrated as scaffolds for revision, reflection, and meaning negotiation rather than as generators of finished text. Students consistently engaged in paraphrasing, tone shifting, and iterative prompt refinement, indicating the development of academic literacy skills that extend beyond surface-level language correction. Importantly, learning gains were observed across prompt strategies and topics, suggesting that educational value arises from task design and learner agency, not from specific AI configurations.

From an instructional perspective, the results highlight the importance of framing AI use around process-oriented outcomes, such as revision depth, semantic preservation, and critical evaluation of AI output. These practices align with sustainability-oriented pedagogies that emphasize autonomy, ethical engagement, and transferable skills. By positioning prompt engineering as a reflective literacy practice rather than a shortcut, the study provides actionable guidance for educators seeking to integrate AI in ways that reinforce—not replace—human judgment and long-term learning capacity.

Primary Contribution of the Study—This study's primary contribution is demonstrating that prompt engineering functions as an iterative academic literacy rather than a simple operational skill in AI-assisted EFL learning. Focusing on computer science students, the study shows how structured prompting, revision, and synthesis enable learners to engage critically with AI-generated text across research summarization, video comprehension, style transformation, and concept mapping tasks. By integrating quantitative analyses (similarity metrics, ANOVA, chi-square tests) with qualitative reflections, the research offers a multidimensional account of how AI-mediated interaction supports language development and academic performance.

The findings reveal that students do not passively accept AI output. Instead, they consistently revise AI-generated text by reorganizing, condensing, and rephrasing content while preserving core meaning. This pattern reflects emerging competencies in academic

literacy, multimodal comprehension, and informed judgment—skills essential for sustainable learning in AI-rich educational environments. Students' ability to filter, adapt, and reinterpret AI-generated information underscores the central role of human agency in effective AI-supported learning.

The study further shows that prompt engineering is most effective when treated as a reflective, iterative process. While structured or role-based prompts improved output clarity, the strongest learning gains occurred when students refined prompts across multiple cycles and actively revised AI responses. These iterative practices fostered deeper semantic processing, metalinguistic awareness, and more academically appropriate expression. Concept mapping tasks provided additional evidence of higher-order learning, as students demonstrated the capacity to integrate AI-generated text with external sources—an early indicator of advanced synthesis and disciplinary knowledge construction.

From a pedagogical perspective, the study advances sustainable education by illustrating how AI tools can enhance learner autonomy, research skills, and academic communication when they support rather than replace human decision-making. Consistent patterns of moderate-to-deep revision across tasks indicate the evolution of responsible AI utilization in accordance with ethical and sustainable learning principles. At the same time, variability in synthesis depth points to the need for explicit instructional guidance, particularly to mitigate over-reliance and support effective multimodal integration.

Taken together, the findings suggest that structured engagement with AI tools can support the development of reflective revision practices and emerging metacognitive awareness among EFL learners, particularly in relation to monitoring meaning, tone, and appropriateness during rewriting tasks. However, these outcomes are best interpreted as task-bounded indicators of metacognitive engagement, rather than as direct evidence of long-term metacognitive growth or sustained language development. The observed patterns of iterative prompting, semantic preservation, and critical revision point to learning practices that are compatible with sustainability-oriented educational frameworks, insofar as they foreground learner agency, responsibility, and purposeful AI use. Future research employing longitudinal designs and independent measures of metacognition will be necessary to determine whether these practices translate into enduring, transferable learning behaviors beyond the immediate instructional context.

In conclusion, this research affirms that AI-assisted prompting and revision can play a transformative role in EFL education, especially in technical and business disciplines that demand both linguistic and domain-specific competence [59]. By positioning prompt engineering as a future-ready literacy grounded in critical reasoning, iterative refinement, and synthesis, the study contributes to a sustainable framework for integrating AI into language education. Future research should examine how these competencies develop longitudinally, transfer across disciplines, and interact with learner agency and ethical awareness as AI continues to reshape higher education.

Author Contributions: Conceptualization, D.R.; methodology, D.R.; data curation, D.R.; writing—original draft preparation, D.R.; writing—review and editing, D.R. and A.S.; project management, G.F.F.; supervision, D.R. and G.F.F.; project administration, D.R., G.F.F. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not require ethics committee approval under the Research Ethics Guidelines of the University of Aizu, as all checklist items for determining the need for review were marked 'No,' and the research involved only anonymized, pre-existing coursework data analyzed without any intervention or identifiable personal information.

Informed Consent Statement: Informed consent from participants was not required because this study qualifies a secondary data analysis or archival study using pre-existing course assignments rather than direct involvement of human subjects. The assignments had already been completed as part of normal coursework, and no new data were collected from students, nor did the study include any interaction or intervention with them. All materials were analyzed in an aggregated and non-identifiable form. Therefore, the use of existing, anonymized documents to examine how students responded to language tasks with and without AI assistance does not necessitate individual informed consent under standard ethical guidelines.

Data Availability Statement: Restrictions apply to the datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Küçükuncular, A.; Ertugan, A. Teaching in the AI era: Sustainable digital education through ethical integration and teacher empowerment. *Sustainability* **2025**, *17*, 7405. [CrossRef]
2. Meilani, Y.; Ardi, A.; Sunarjo, R.A.; Berlianto, M.; Meyvanali, M. Human Technology Sustainability: Leveraging Artificial Intelligence in Education for Inclusive and Ethical Futures. 2025. Available online: https://ssrn.com/abstract=5296904 (accessed on 5 January 2026).
3. Lin, C.C.; Huang, A.Y.Q.; Lu, O.H.T. Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learn. Environ.* **2023**, *10*, 41. [CrossRef]
4. Kramsch, C. *Language and Culture*; Oxford University Press: Oxford, UK, 2002.
5. Ushioda, E. Researching L2 Motivation: Past, Present, and Future. In *The Palgrave Handbook of Motivation for Language Learning*; Lamb, M., Csizér, K., Henry, A., Ryan, S., Eds.; Palgrave Macmillan: Cham, Switzerland, 2019. [CrossRef]
6. Van Lier, L. *The Ecology and Semiotics of Language Learning: A Sociocultural Perspective*; Springer: Boston, MA, USA, 2004. [CrossRef]
7. Li, X. AI, learner identity, and sustainable language development. *Appl. Linguist. Rev.* **2024**, *15*, 77–101.
8. Stickler, U.; Hampel, R. Transforming Teaching: New Skills for Online Language Learning Spaces. In *Developing Online Language Teaching. New Language Learning and Teaching Environments*; Hampel, R., Stickler, U., Eds.; Palgrave Macmillan: London, UK, 2015. [CrossRef]
9. Ushioda, E. *Motivation, Language Identity, and the L2 Self*; Multilingual Matters: Bristol, UK, 2011.
10. Strielkowski, W.; Grebennikova, V.; Lisovskiy, A.; Rakhimova, G.; Vasileva, T. AI-driven adaptive learning for sustainable educational transformation. *Sustain. Dev.* **2025**, *33*, 1921–1947. [CrossRef]
11. Federiakin, D.; Molerov, D.; Zlatkin-Troitschanskaia, O.; Maur, A. Prompt engineering as a new 21st-century skill. *Front. Educ.* **2024**, *9*, 1366434. [CrossRef]
12. Lee, D.; Palmer, E. Prompt engineering in higher education: A systematic review to help inform curricula. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 7. [CrossRef]
13. Wulyani, A.N.; Widiati, U.; Muniroh, S.; Rachmadhany, C.D.; Nurlaila, N.; Hanifiyah, L.; Tengku Sharif, T.I.S. Patterns of Utilizing AI-Assisted Tools among EFL Students: Need Surveys for Assessment Model Development. *J. Lang. Lang. Teach.* **2024**, *27*, 561–565. [CrossRef]
14. Kohnke, L.; Zou, D.; Zhang, R. Prompting ChatGPT in EAP and ESP contexts: A typology of learner-generated prompts. *J. Engl. Acad. Purp.* **2023**, *62*, 101239.
15. Lee, J.; Lee, H. AI-assisted language learning and sustainable digital literacies in higher education. *Educ. Technol. Soc.* **2024**, *27*, 45–59.
16. Woo, D.J.; Guo, K.; Susanto, H. Exploring EFL students' prompt engineering in human–AI story writing: An activity theory perspective. *Interact. Learn. Environ.* **2025**, *33*, 863–882. [CrossRef]
17. Woo, D.J.; Guo, K.; Susanto, H. Exploring EFL students' prompt engineering in human–AI story writing. *arXiv* **2023**, arXiv:2306.01798. [CrossRef]
18. Kim, J.; Yu, S.; Lee, S.S.; Detrick, R. Students' prompt patterns and Its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *J. Res. Technol. Educ.* **2025**, 1–18. [CrossRef]
19. Sari, D. Iterative feedback and the development of rhetorical awareness in L2 writing. *J. Writ. Res.* **2023**, *15*, 349–371.
20. Mzwri, K.; Turcsányi-Szabo, M. The Impact of Prompt Engineering and a Generative AI-Driven Tool on Autonomous Learning: A Case Study. *Educ. Sci.* **2025**, *15*, 199. [CrossRef]
21. Bhatti, A. Strategic Prompt Engineering for Enhancing AI-Generated Content in English Language Teaching: Empowering EFL Contexts. *Int. J. Comput. Assist. Lang. Learn. Teach.* **2026**, *16*, 1–30. [CrossRef]

22. Miyazoe, T. Effective Prompts for EFL Writers: Leveraging ChatGPT for Writing Aid. In *The IAFOR International Conference on Education—Hawaii (IICE 2025) Official Conference Proceedings*; IAFOR: Nagoya, Japan, 2025; pp. 137–145. [CrossRef]
23. Sudta, K.; Siridej, N. ESL/EFL Teachers' Use of Prompt Engineering for AI-Assisted Material Design: Opportunities and Concerns. 2024. Available online: https://has.hcu.ac.th (accessed on 12 December 2025).
24. Freeman, J. UK Universities Warned to "Stress-Test" Assessments as 92% of Students Use AI. 2025. Available online: https://www.theguardian.com/education/2025/feb/26/uk-universities-warned-to-stress-test-assessments-as-92-of-students-use-ai (accessed on 15 January 2026).
25. Student Generative AI Survey 2025. February 2025. Available online: https://www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf (accessed on 9 January 2026).
26. Killemsetty, N. Are Indian Classrooms Ready for the AI Leap? 2025. Available online: https://360info.org/are-indian-classrooms-ready-for-the-ai-leap/ (accessed on 19 January 2026).
27. PTI. AI Adoption Grows in Indian B-Schools, but Only 7% Faculty Are Expert Users: Survey. 2025. Available online: https://economictimes.indiatimes.com/industry/services/education/ai-adoption-grows-in-indian-b-schools-but-only-7-per-cent-faculty-are-expert-users-survey/articleshow/124192323.cms (accessed on 19 January 2026).
28. Davey, E. University Assessments to Be Reviewed as 92% of Students Use AI. 2025. Available online: https://theboar.org/2025/03/university-assessments-to-be-reviewed-as-92-of-students-use-ai/ (accessed on 9 January 2026).
29. Engageli. 20 Statistics on AI in Education to Guide Your Learning. 2025. Available online: https://www.engageli.com/blog/ai-in-education-statistics (accessed on 8 January 2026).
30. Thomann, H. Scaffolding through prompts in digital learning: Meta-analytic evidence. *Comput. Educ.* **2025**, *200*, 104–127.
31. Favero, L.; Pérez-Ortiz, J.A.; Käser, T.; Oliver, N. Enhancing Critical Thinking in Education by Means of a Socratic Chatbot. *arXiv*. 2024. [CrossRef]
32. Qian, Y. Prompt engineering in education: A systematic review of approaches and educational applications. *J. Educ. Comput. Res.* **2025**, *63*, 1782–1818. [CrossRef]
33. Ng, L.; Singh, S.K.; Ang, E.T.; Jumat, N. *Scaffolding Higher-Order Thinking Through AI Chatbots: A Multi-Domain Study*; Singapore Institute of Technology: Singapore, 2025. Available online: https://irr.singaporetech.edu.sg/articles/conference_contribution/Scaffolding_Higher-Order_Thinking_Through_AI_Chatbots_A_Multi-Domain_Study/29231549/1/files/56278418.pdf (accessed on 19 January 2026).
34. Liu, S.; Guo, X.; Hu, X.; Zhao, X. Advancing generative intelligent tutoring systems: Socratic-style prompts for adaptive reasoning. *Electronics* **2024**, *13*, 4876.
35. Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv* **2022**, arXiv:2205.10625. [CrossRef]
36. Post, N.Y. Majority of Students See Responsible AI Use as Key to Career Success, New Research Says. 2025. Available online: https://nypost.com/2025/07/29/tech/majority-of-students-see-responsible-ai-use-as-key-to-career-success-new-research-says/ (accessed on 5 January 2026).
37. El-Dakhs, D. L2 paraphrasing and meaning preservation: Insights from AI-supported rewriting tasks. *System* **2024**, *117*, 103133.
38. Mazgutova, D.; McCray, G. An exploratory analysis of revision behavior development of L2 writers on an intensive English for academic purposes program using Bayesian methods. *Front. Commun.* **2023**, *7*, 934583. [CrossRef]
39. Peltzer, K.; Lira Lorca, A.; Krause, U.M.; Graham, S.; Busse, V. Effects of feedback on deep-level features of argumentative writing over multiple drafts: Insights from an intervention study with secondary EFL students. *Read. Writ.* **2025**, *39*, 333–364. [CrossRef]
40. Liu, M.; Zhang, T. Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Comput. Educ.* **2024**, *211*, 104977. [CrossRef]
41. Smutny, P.; Schreiberova, P. ChatGPT and academic writing accuracy: Impacts on coherence, precision, and disciplinary fit. *Comput. Educ. Artif. Intell.* **2023**, *5*, 100151.
42. Zhang, Y. Semantic similarity metrics as indicators of L2 writing development. *Assess. Writ.* **2023**, *58*, 100679.
43. Nguyen, T.; Pham, A. Domain-specific reasoning with generative AI: Implications for higher education. *IEEE Trans. Learn. Technol.* **2024**, *17*, 88–102.
44. Wang, Y. Student perceptions of ChatGPT in EFL academic tasks. *Educ. Technol. Soc.* **2024**, *27*, 45–58.
45. Hsu, H.-Y.; Chen, Y. Multimodal comprehension and the use of AI tools in EFL academic tasks. *Educ. Technol. Res. Dev.* **2023**, *71*, 1543–1565.
46. Wang, K.; Cui, W. Artificial intelligence in higher education: The impact of need satisfaction on AI literacy mediated by self-regulated learning strategies. *Behav. Sci.* **2025**, *15*, 165. [CrossRef]
47. Chee, H.; Ahn, S.; Lee, J. A competency framework for AI literacy. *Br. J. Educ. Technol.* **2024**, *56*, 2146–2182. [CrossRef]
48. Fuchs, C. Limitations of ChatGPT in EFL contexts. *TESOL Q.* **2023**, *57*, 1520–1538.
49. Chen, A.; Zhang, Y.; Jia, J.; Liang, M.; Cha, Y.; Lim, C.P. A Systematic Review and Meta-Analysis of AI-Enabled Assessment in Language Learning: Design, Implementation, and Effectiveness. *J. Comput. Assist. Learn.* **2025**, *41*, e13064. [CrossRef]

50. Taguchi, N. Pragmatic socialization in an English-medium university in Japan. *Int. Rev. Appl. Linguist. Lang. Teach.* **2014**, *52*, 157–182. [CrossRef]
51. Taguchi, N. Second language acquisition and pragmatics. In *The Routledge Handbook of Second Language Acquisition and Pragmatics*; Taguchi, N., Ed.; Routledge: New York, NY, USA, 2019; pp. 1–14.
52. Rose, H.; Galloway, N. *Global Englishes for Language Teaching*; Cambridge University Press: Cambridge, UK, 2019. [CrossRef]
53. Holmes, W.; Bialik, M.; Fadel, C. *Artificial Intelligence in Education: Promises and Implications*; Center for Curriculum Redesign: Boston, MA, USA, 2019; Available online: https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf (accessed on 5 January 2026).
54. Miao, F.; Holmes, W.; Huang, R.; Zhang, H. *AI and Education: Guidance for Policymakers*; UNESCO: Paris, France, 2021.
55. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
56. Devatine, N.; Abraham, L. Assessing human editing effort on LLM-generated texts via compression-based edit distance. *arXiv* **2024**, arXiv:2412.17321. [CrossRef]
57. Ma, Y.; Qing, L.; Liu, J.; Kang, Y.; Zhang, Y.; Lu, W.; Liu, X.; Cheng, Q. From model-centered to human-centered: Revision Distance as a metric for Text Evaluation in LLMs-based Applications. *arXiv* **2024**, arXiv:2404.07108. [CrossRef]
58. Cahyani, A.D.; Fathoni, M.; Rachman, F.H.; Basuki, A.; Amin, S.; Khotimah, B.K. Automatic essay scoring: Leveraging Jaccard coefficient and Cosine similarity with n-gram variation in vector space model approach. *arXiv* **2025**, arXiv:2510.15311. [CrossRef]
59. Roy, D. Pedagogical restructuring of business communication courses: AI-enhanced prompt engineering in an EFL teaching context. In *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*; Ilic, P., Casebourne, I., Wegerif, R., Eds.; In Intelligent systems reference library; Springer: Cham, Switzerland, 2024; Volume 261, pp. 247–287. [CrossRef]