BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies
https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Tracking Editorial Footprints: A Process-Oriented Analysis of ChatGPT Reliance in Student Writing

Jungmin Kwon [1] , ***Youngsun Lee*** [2*]

[1] *Department of Educational Technology, Seoul National University of Education, Seoul 06639, Republic of Korea*

[2] *Department of Special Education, Ewha Womans University, Seoul 03760, Republic of Korea*

ABSTRACT

In this study, we propose a process-oriented framework centered on "editorial footprints," which we define as the observable steps in a writer's drafting and revision process when using generative AI. Fifteen female undergraduate students completed two writing tasks using ChatGPT: one under a quick, minimal-effort condition and another under a thorough, high-effort condition. Participants edited a shared rough draft in Google Docs, while their entire interactions with ChatGPT were recorded and qualitatively analyzed. Results show that while the final text lengths were similar, students in the thorough condition made significantly more edits and employed a broader range of ChatGPT prompts, producing work with greater depth, logical coherence, and style consistency, which left more editorial footprints throughout the writing process. These findings reveal distinct patterns of engagement, prompting, and revision between the two conditions and demonstrate the limitations of current AI detectors, which overlook the full scope of the writing process. Our discussion emphasizes that detection of AI-generated writing should incorporate analysis of the writer's interaction histories and revision behaviors with generative AI tools. We further suggest that understanding these process-based indicators is essential not only for distinguishing AI-assisted writing but also for fostering educational practices that encourage meaningful, reflective engagement with AI in writing.

Keywords: Writing; Higher Education; ChatGPT; LLM; AI Detection; Prompts

*CORRESPONDING AUTHOR:
Youngsun Lee, Department of Special Education, Ewha Womans University, Seoul 03760, Republic of Korea; Email: ylee@ewha.ac.kr

ARTICLE INFO
Received: 12 May 2025 | Revised: 3 June 2025 | Accepted: 9 June 2025 | Published Online: 14 June 2025
DOI: https://doi.org/10.30564/fls.v7i6.9985

CITATION
Kwon, J., Lee, Y., 2025. Tracking Editorial Footprints: A Process-Oriented Analysis of ChatGPT Reliance in Student Writing. Forum for Linguistic Studies. 7(6): 941–954. DOI: https://doi.org/10.30564/fls.v7i6.9985

1. Introduction

Since the emergence of ChatGPT in late 2022, educational institutions have attempted to ban its use citing concerns about academic dishonesty [1]. For example, the New York City Department of Education restricted access to ChatGPT on school devices and networks due to fears that students could exploit AI to bypass authentic learning and cheat on assignments [2]. Likewise, several universities around the world have begun revising their academic integrity policies to address AI-generated content more explicitly [3,4], while others have strengthened their plagiarism detection protocols by incorporating AI-detection software [5].

In response, researchers and developers have created AI-based "ChatGPT detectors" to distinguish human-written text from AI-generated text [6-8]. For example, GPTZero is a tool that analyzes text for linguistic markers such as syntactic complexity and "burstiness" to estimate whether a passage was written by a human or generated by ChatGPT [9]. Other platforms, such as Turnitin's AI-writing detection solution, apply machine learning algorithms to identify potential instances of AI-generated prose by scrutinizing semantic coherence and stylistic irregularities [10].

Although no formal report exists on the percentage of lecturers and professors using AI detectors in higher education, a 2024 survey in the United States reported that 68% of K–12 teachers use these tools as they struggle with concerns regarding students who use generative AI to cheat on assignments [11]. While these detection tools represent a significant step toward mitigating academic dishonesty, critics note that the detectors remain susceptible to false positives and false negatives [12]. For example, texts that demonstrate strong logical organization and high linguistic quality risk being flagged as AI-generated [13]. Recent studies concur that identifying AI-generated text is inherently fallible and no current detection method can ensure complete accuracy [14,15]. This issue is especially concerning in academic environments where the stakes of a misclassification can be profound. From a student rights perspective, these false positives, in which genuinely human-produced work is labeled as AI-generated, can lead to unjust repercussions that range from lowered grades to formal disciplinary actions [6]. Although many institutions have turned to AI detectors to maintain academic integrity, the potential harm inflicted by incorrect judgments underscores the necessity of more nuanced, reliable detection practices that reduce the unintended consequences of these emerging technologies.

Current detectors focus solely on analyzing a piece of writing after it is already complete. Writing, however, is not merely a static end product. Numerous scholars have argued that writing is a complex, iterative activity that extends far beyond the final artifact [16–18]. Writing is a process that includes stages of brainstorming, drafting, revising, and reflecting, each of which shapes the developing text in distinct ways [19]. Cognitive process theories of writing emphasize the significance of monitoring a piece of text through multiple edits since such revisions shed light on the writer's thought processes, problem-solving strategies, and degree of engagement [17,20].

The choices that writers make during composition, as well as the nature and frequency of their revisions, can provide valuable insights into their underlying cognitive flow and motivations [18]. These are called "editorial footprints" [17]. Editorial footprints, or digital fingerprints as described by Salman and Alexandron [21], refer to the traceable record of a writer's iterative process of planning, drafting, and revising, which reveals their cognitive flow and engagement beyond the final text. The motivation for this research stems from the fact that purely outcome-focused detectors cannot capture these micro-level processes. Unlike current tools that analyze only a single, static version of a document, a process-oriented detector could enable educators and researchers to identify unique markers of human authorship by examining how a writer conceptualizes ideas, reorganizes content, and refines language over time [16,19]. This focus on process is a central tenet of modern composition theory and underscores the limitations of treating writing as an isolated, one-step task [22]. By tracking the writing process, researchers may better determine whether AI tools simply aided a student's cognitive workflow or fully replaced it. This methodological shift, we believe, holds promise for creating more accurate and ethical approaches to detecting AI usage.

To assess the feasibility of a process-based approach to AI detection, we examined editorial footprints by tracking a writer's iterative process. Our goal was to understand how varying levels of ChatGPT reliance affect students' writing processes, editorial behaviors, and final outputs. Specifically, we focused on two writing conditions: one in which students rely almost entirely on ChatGPT (the Quick Writing Condition), and another in which they use ChatGPT primarily for partial assistance (the Thorough Writing Condition). We compared these two extremes (quick vs. thorough) because they represent distinct writing processes—minimal interaction versus iterative refinement. Identifying measurable differences in each process could shed light on features that a process-based detection method can leverage.

Our overarching research question was as follows: How do students behave differently in the writing process when they use ChatGPT for quick completion versus when they engage in thorough, high-quality writing? To address this question, we posed three sub-questions:

How does writing quality differ when students use ChatGPT primarily for quick task completion versus when they seek to produce a thorough, high-quality writing assignment?

How do edit counts and types differ when students use ChatGPT primarily for quick task completion versus when they aim to create a thorough, high-quality writing assignment?

How do prompt counts and types differ when students use ChatGPT primarily for quick task completion versus when they strive to craft a thorough, high-quality writing assignment?

2. Methods

2.1. Participants

Fifteen female undergraduate students from two universities participated in this study; they majored in various fields of education, including Special Education, Early Childhood Education, Elementary Education, Computer Education, History Education, and Lifelong Education. Their ages ranged from 19 to 23 years, with a mean age of 21.4 years (SD = 1.24). These students, enrolled at various academic levels from freshman to senior year, participated voluntarily. They were asked to self-report their writing ability on a scale from 1 (low) to 3 (high), yielding a mean score of 2.07 (SD = 0.70). Additionally, they were asked to rate their experience using ChatGPT on a scale from 1 (never) to 3 (frequently), with a mean score of 2.2 (SD = 0.77).

2.2. Research design

To control for individual variability, the study employed a within-subject experimental design examining how students utilize ChatGPT and characterize their writing and editing processes under different conditions. Fifteen students were asked to complete two writing tasks: a first task requiring low effort (Quick Writing condition), and a second task requiring high effort (Thorough Writing condition). Each participant completed both tasks in the specified order, allowing for the comparison of individual differences in ChatGPT use and writing behavior across these two conditions. This design enabled the examination of participants' performance changes and behavior across both conditions.

2.3. Assignment

Prior to the writing tasks, participants received specific instructions. For the first task (Quick Writing), they were told, "For the first writing task, your primary goal is to complete the assignment. Imagine yourself as a student who is not too excited about the assignment and just needs to submit one." Upon completion of the first task, the second task (Thorough Writing) was administered, with instructions to "For the second writing task, your primary goal is to write a good essay. Imagine you are a student wanting to receive a good grade on this assignment." In both tasks, participants were informed they could use ChatGPT freely, according to their needs.

To maintain consistency in genre, style, and quality, participants were provided with a one-page rough draft created by the researchers to use as a starting point. This approach served two purposes. First, tracking revisions in standard word processors (e.g., MS Word, Google Docs) is not possible when beginning with a blank document, as all edits are lumped into a single time edit. However, if participants edited a document initiated by someone else, more detailed editorial footprints were recorded—particularly in Google Docs. Second, providing a uniform rough draft allowed for control over content and difficulty, ensuring that each participant worked on an assignment of comparable scope and complexity. No restrictions were imposed on text length or the time allocated for revisions. All participants received the same draft under both conditions, and the order of tasks was kept consistent to minimize potential learning effects. The quick writing task was presented first, since participants were expected to spend less time and exert less cognitive effort. This was followed by the thorough writing task, where participants invested greater time and engaged in deeper cognitive processing.

2.4. Data Collection

Data were collected using two online platforms. First, we used Google Docs to observe participants' writing behaviors and collect writing data. Participants were given a 450-word rough draft, which they had to use as the basis for their assignment. Each participant agreed to complete the assignment exclusively within the provided document, with revision history enabled to monitor the writing process. Second, to examine the ways participants utilized ChatGPT, each participant was given a unique ChatGPT account and password. Their ChatGPT conversation history was automatically recorded. Upon notification of task completion by the participants, the researchers accessed the ChatGPT accounts, exported the content as a text file, and additionally captured all usage data through screenshots for backup purposes. This procedure was repeated for the second task.

2.5. Measurement and Analysis of Writing Behavior and Quality

To measure the quality of the writings, we evaluated text length, format quality, and content quality. For text length, we counted the number of letters, words, and paragraphs, which were directly available from Google Docs. For format quality, we examined style (coherent or mixed), modifications to the title, inclusion of the author's name, and the presence of any observable hallucinations. For content quality, we assessed student writing using three criteria: depth of content, logical coherence, and use of references or specific examples. Two independent raters conducted blind evaluations, and the average of their scores was used as the format and content quality score.

To analyze writing behavior, we investigated both the frequency and types of edits. For the number of edits, we counted how many times participants revised their text. For the types of edits, we tracked participants' writing using Google Docs' revision history.

To measure final-writing quality, we focused on text length, format quality, and content quality. Text length was recorded using the word processor's automatic count. Format quality was evaluated by examining style (coherent or mixed), modifications to the title, inclusion of the author's name, and any observable hallucinations. Content quality was assessed on three criteria: content depth, logical coherence, and the use of references or specific examples. Two independent raters conducted blind assessments, and the average of their ratings was used as the final score for both format and content quality.

2.6. Measurement and Analysis of ChatGPT Use

To understand how participants used ChatGPT for the two assignments, we analyzed participants' interactions with ChatGPT using Atlas.ti, a qualitative research software. The researchers manually coded the data to categorize the types of prompts participants used (e.g., "Revise title," "Write conclusion," "Give examples"). After initial coding, these categories were refined into second-level codes, which formed the basis for subsequent analysis and interpretation.

For the visual representation, we assigned edge weights to indicate how frequently each code appeared in the dataset. Specifically, we tallied the number of times each code was referenced across participant data. This weighted approach not only revealed which types of prompts were utilized in writing but also illustrated the strength of these associations, offering a more nuanced view of the data's underlying structure.

2.7. Reliability

For all quantitative measures, including edit counts and types, student writings were independently evaluated by two authors, and inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC). The ICC for edit count, types of edits and their count were 1.00; and text length and paragraph count were 0.99. Regarding format quality, the ICC for paragraphing was 0.80, and for writing style, inclusion of the author's name, and presence of noticeable hallucinations, it was 1.00. For word count, the word processor automatically calculated character and word counts, so no additional reliability checks were needed. To assess the quality of writing, two public school teachers who were not involved in this study were asked to score the texts. The ICC for content depth was 0.62, logical coherence was 0.81, and ICC for use of references or concrete examples was 1.0. To classify the types of questions in the ChatGPT usage records, the first author initially performed 100% coding. The resulting code list was then given to the second author, who independently coded 100% of the data using the provided codes to examine inter-rater reliability. Inter-rater reliability was 99.98%.

3. Results

3.1. Research Question 1: How Does Students' Writing Quality Differ When They Use ChatGPT Primarily for Quick Task Completion versus When They Aim to Craft a Thorough, High-Quality Writing Assignment?

A Shapiro–Wilk test was used to assess the normality of difference scores for each measured variable. As shown in **Table 1**, Letter Count, Word Count, Paragraph Count, Style Consistency, Paragraphing Error, Content Depth, and Logical Coherence did not significantly differ from a normal distribution (*p* > 0.05). Therefore, these variables were analyzed using paired t-tests. In contrast, Edit Count, Hallucination, Editing Title, Author's Name, and Use of Reference showed statistically significant deviations from normality (*p* < 0.05). These variables were therefore analyzed using Wilcoxon signed-rank tests.

Table 1. Normality Test for Measured Variables(Shapiro-Wilk).

Measured Variables	W	*p*	Paired t-Test	Wilcoxin Signed-Rank
Letter Count	0.941	0.392		V
Word Count	0.947	0.482		V
Paragraph Count	0.909	0.131		V
Style Consistency	0.776	0.002	V	
Paragraphing Error	0.899	0.091		V
Hallucination	0.284*	< 0.001	V	
Editing Title	0.413*	<0.001	V	
Author's Name	0.284*	< 0.001	V	
Use of Reference	0.743*	<0.001	V	
Content Depth	0.877	0.043	V	
Logical Coherence	0.916	0.169		V

**p* < 0.05.

3.1.1. Text Length

Results in **Table 2** indicate that no significant difference can be found in either Letter Count (t(14) = 0.314, *p* = 0.758), or Word Count (t(14) = 0.31, *p* = 0.761). Paragraph Count also did not reach statistical significance (t(14) = -1.801, *p* = 0.093). This suggests that, regardless of whether students wrote the paper quickly or thoroughly, final length of their writing was not affected.

Table 2. Quick vs Thorough Edit Counts and Text Length.

	n	Quick		Thorough		*df*	*t*	*p*	Effect Size
		M	*SD*	*M*	*SD*				
Letter Count	15	2770.33	937.80	2866.20	599.50	34	0.314	0.758	0.081
Word Count	15	666.27	225.96	688.47	143.20	14	0.31	0.761	0.080
Paragraph Count	15	10.60	5.15	8.00	2.45	14	-1.801	0.093	-0.465

*p < 0.05.

3.1.2. Writing Quality

We compared the quality of students' written output in terms of both format and content (**Table 3**). With respect to format quality, style consistency was significantly higher for thorough writing (M = 1.00, SD = 0.00) than for quick writing (M = 0.067, SD = 0.489), W = 2.646, p = 0.019, indicating that students were more consistent in their overall style when they invested additional effort in revising. Interestingly, paragraphing errors were more frequent in the thorough condition (M = 0.60, SD = 1.056) compared to the quick condition (M = 0.20, SD = 0.414), W = −9.320, p < 0.001, suggesting that although students worked to improve style consistency, they occasionally failed to paragraph effectively. No significant differences emerged for hallucination errors, edited titles, or whether students added the author's name(p > 0.05).

Table 3. Format and Content Quality.

	n	Quick		Thorough		*dF*	Statistics	*p*	Effect Size
		M	SD	M	SD				
Format Quality								.	
Style Consistency	15	0.067	0.489	1.00	0.00	14	2.646*	0.019	0.683
Paragraphing Error	15	0.2	0.414	0.6	1.056	14	-9.320*	<0.001	-2.410
Hallucination	15	0.067	0.258	0	0		0	1	-1.000
Edited Title	15	0.067	0.258	0.267	0.458		6	0.149	1.000
Put Author's Name	15	0	0	0.067	0.258		1	1	1.000
Content Quality									
Use of Reference	15	0	0	0.533	1.356		6	0.181	1
Content Depth	15	3.667	1.447	4.933	0.961	14	2.801*	0.014	0.723
Logical Coherence	15	3.733	1.387	4.733	1.033	14	2.185*	0.046	0.564

*p < 0.05.

3.1.3. Content Quality

As for content quality, there were no significant differences in the use of references (p = 0.181). However, content depth was significantly greater in thorough assignments (M = 4.933, SD = 0.961) compared to quick tasks (M = 3.667, SD = 1.447); t(14) = 2.801, p = 0.014, reflecting a more substantive exploration of the topic. A similar pattern emerged for logical coherence, which was significantly higher in thorough assignments (M = 4.733, SD = 1.033) than in quick ones (M = 3.733, SD = 1.387); t(14) = 2.185, p = 0.046. These findings suggest that students producing thorough writing engaged in more depth of analysis and maintained stronger overall coherence, while also applying a more consistent style relative to those focusing on quick task completion.

3.2. Research Question 2: How Do Edit Counts and Types Differ When Students Use ChatGPT Primarily for Quick Task Completion versus When They Aim to Craft a Thorough, High-quality Writing Assignment?

A Shapiro–Wilk test was first performed to check normality for each variable. If the normality test was significant (p < 0.05), a Wilcoxon signed-rank test was used; otherwise, a paired t-test was used (**Table 4**). Wilcoxon signed-rank test was used for statistical testing for all variables except “Delete space.”

Table 4. Normality Test for Each Measures(Shapiro-Wilk).

Measured Variables	W	*p*	Paired t-Test	Wilcoxon Signed-Rank
Total edit count	0.742	<0.001		V
Delete letter or word	0.872	0.036		V
Add letter or word	0.729	<0.001		V
Delete paragraph	0.844	0.014		V
Add paragraph	0.826	0.008		V
Delete space	0.921	0.202	V	
Add space	0.868	0.032		V
Change, edit	0.541	<0.001		V
Switch	0.284	<0.001		V
Formatting	0.672	<0.001		V
Footnote	0.284	<0.001		v

*p < 0.05.

3.2.1. Edit Count

A Wilcoxon signed-rank test indicated that the Thorough condition (M = 27.4, SD = 25.41) had a significantly higher total edit count than the Quick condition (M = 9.6, SD = 9.91), p = 0.013, effect size = 0.736 (**Table 5**).

Table 5. Edit Counts and Type of Edits.

	n	Quick		Thorough		df	Statistic	*p*	Effect Size
		M	SD	M	SD				
Total edit count	15	9.6	9.912	27.4	25.41		2.849*	0.013	0.736

Table5, *Cont.*

	n	Quick		Thorough		df	Statistic	*p*	Effect Size
		M	SD	M	SD				
Types of edits									
Delete word or sentence	15	1.667	3.457	3.4	2.971		11.00*[a]	0.016	-0.758
Add word or sentence	15	2.733	2.939	8.8	9.98		18*	0.018	-0.7
Delete paragraph	15	0.8	0	0.467	0.64		26.50[b]	0.673	0.178
Add paragraph	15	0.267	0.594	0.733	1.223		4.50[d]	0.236	-0.571
Delete space	15	0.2	0.561	1	1.069	14	-2.256*	0.041	-0.583
Add space	15	0.467	1.06	1.467	2.031		13.00[e]	0.077	-0.606
Change, edit	15	3.2	4.427	10.333	12.993		0.00*[a]	0.002	-1
Switch	15	0	0	0.0667	0.258		0.00[f]	1	-1
Formatting	15	0.267	0.799	0.8667	2.167		2.50[g]	0.461	-0.5
Footnote	15	0	0	0.2667	1.033		0.00[f]	1	-1

*p < 0.05.
[a] 2 pair(s) of values were tied.
[b] 6 pair(s) of values were tied.
[d] 9 pair(s) of values were tied.
[e] 4 pair(s) of values were tied.
[f] 14 pair(s) of values were tied.
[g] 11 pair(s) of values were tied.

3.2.2. Edit Types

A Wilcoxon signed-rank test showed participants in the Thorough condition (M = 3.4, SD = 2.97) made more “Delete letter or word” edits than those in the Quick condition (M = 1.67, SD = 3.46), *p* = 0.016, effect size = –0.758; made significantly more “Add letter or word “edits in the Thorough condition (M = 8.8, SD = 9.98) compared to the Quick condition (M = 2.73, SD = 2.94), *p* = 0.018, effect size = –0.700; and made significantly higher frequency of “Change/edit” actions in the Thorough condition (M = 10.33, SD = 12.99) compared to the Quick condition (M = 3.20, SD = 4.43), *p* = 0.002, effect size = –1.00. A paired t-test revealed that participants in the Thorough condition (M = 1.0, SD = 1.07) were more likely to “Delete space” than those in the Quick condition (M = 0.2, SD = 0.56), t(14) = –2.256, *p* = 0.041, effect size = –0.583. No significant differences emerged between the two conditions for “Delete paragraph”, “Add paragraph”, “Add space”, “Switch”, “Formatting”, or “Footnote” edits (**Table 5**).

Figure 1 is a visual representation of the statistics. Participants in the Thorough condition produced more edits across most categories than those in the Quick condition. The largest discrepancies appeared in Add Word or Sentence and Change, while moderate differences were observed in Delete Word or Sentence and Delete Space. This suggests that students who used ChatGPT more constructively engaged in a more iterative and detailed revision process, therefore leaving more editorial footprints.

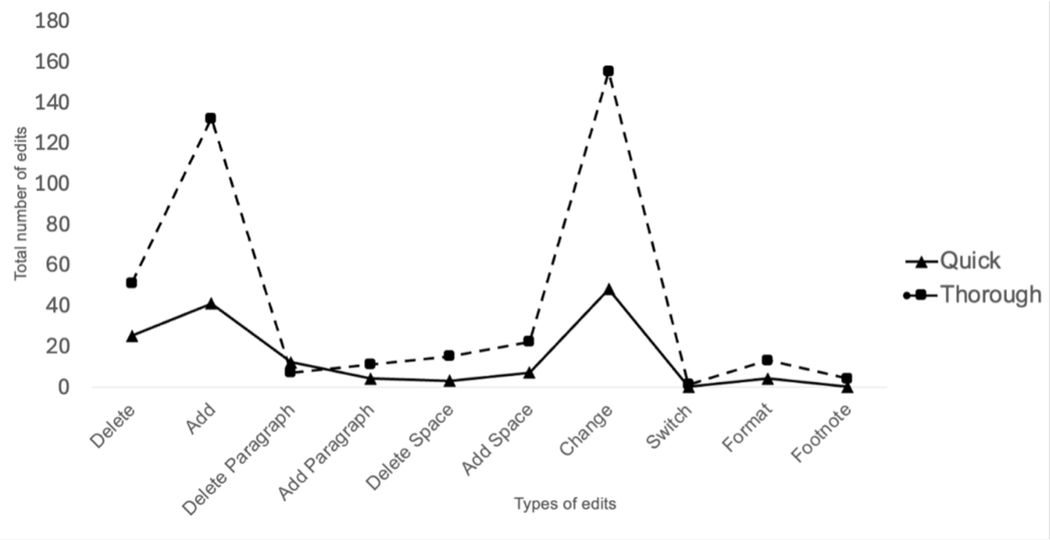

Figure 1. Total number of edits by type.

3.3. Research Question 3: How Do Prompt Counts and Types Differ When Students Use ChatGPT Primarily for Quick Task Completion versus When They Aim to Craft a Thorough, High-Quality Writing Assignment?

For research question 3, we conducted a visual analysis to understand how students prompted in two conditions. In **Figures 2** and **3**, arrow thickness represents how frequently students used each prompt type. For example, if the count for "Increase volume" was 15, the arrow thickness was configured as 15 pts. By looking at which prompt categories have the thickest arrows in each diagram, we can identify the most common uses of ChatGPT for each condition. In the Quick condition students issued a total of 92 prompts, whereas those in the Thorough condition used more than twice as many (211 prompts in total).

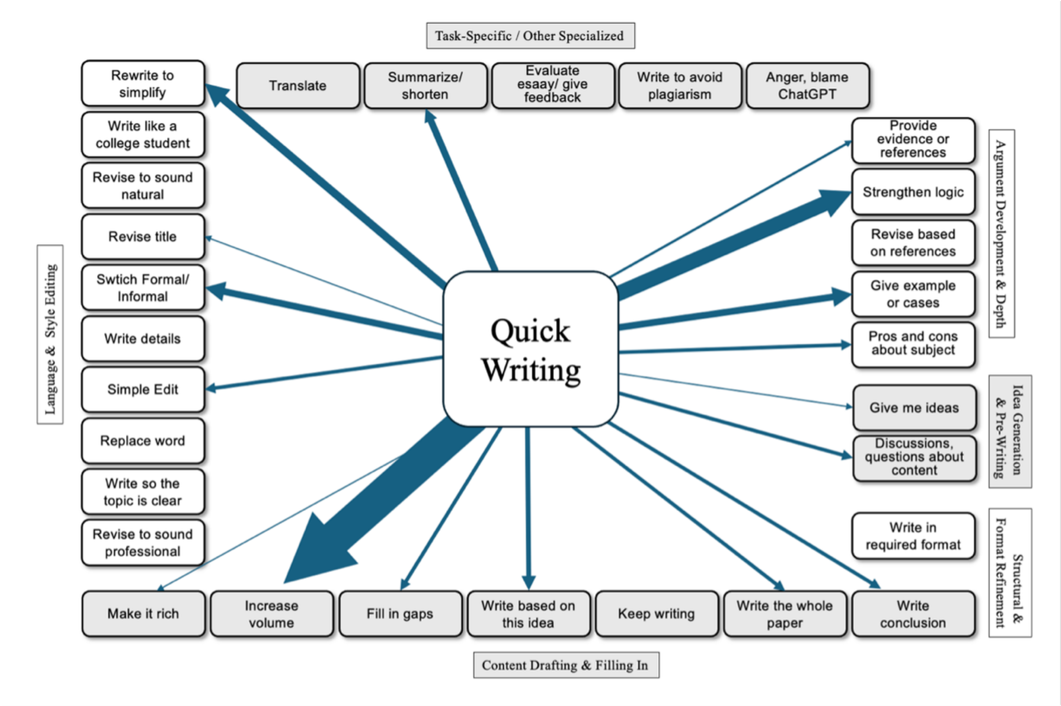
Task-Specific / Other Specialized
Translate
Summarize/ shorten
Evaluate esaay/ give feedback
Write to avoid plagiarism
Anger, blame ChatGPT
Language & Style Editing
Rewrite to simplify
Write like a college student
Revise to sound natural
Revise title
Swtich Formal/ Informal
Write details
Simple Edit
Replace word
Write so the topic is clear
Revise to sound professional
Quick Writing
Argument Development & Depth
Provide evidence or references
Strengthen logic
Revise based on references
Give example or cases
Pros and cons about subject
Idea Generation & Pre-Writing
Give me ideas
Discussions, questions about content
Structural & Format Refinement
Write in required format
Write conclusion
Make it rich
Increase volume
Fill in gaps
Write based on this idea
Keep writing
Write the whole paper
Content Drafting & Filling In

Figure 2. Types of ChatGPT Prompts in Quick Writing Conditions.

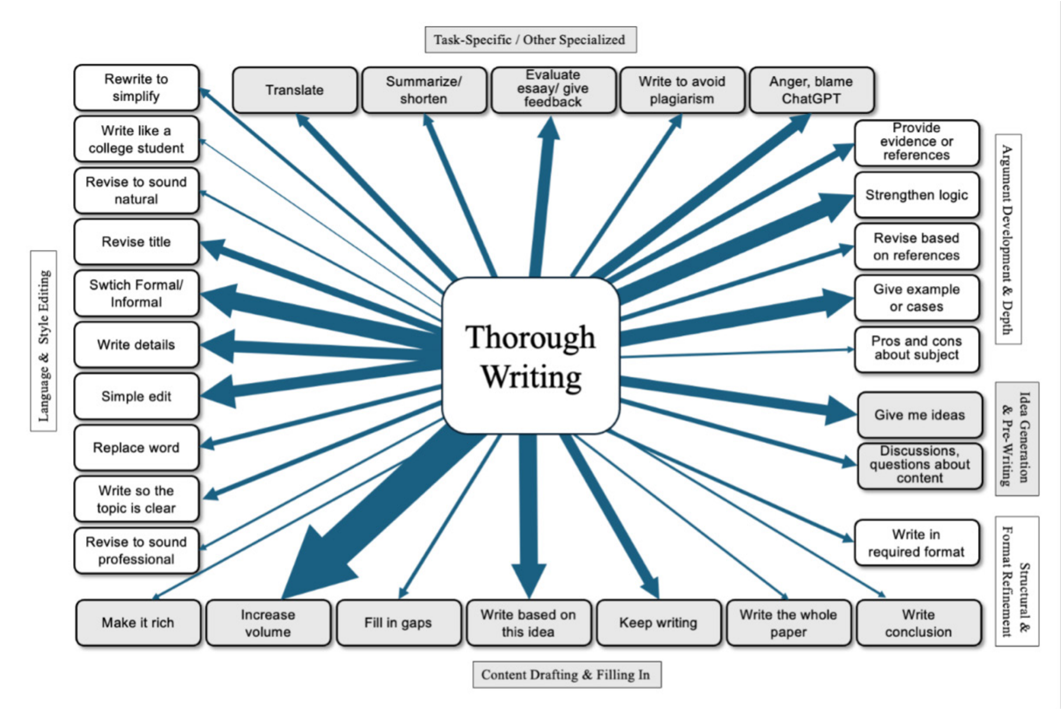
Task-Specific / Other Specialized
Rewrite to simplify
Translate
Summarize/ shorten
Evaluate esaay/ give feedback
Write to avoid plagiarism
Anger, blame ChatGPT
Write like a college student
Provide evidence or references
Revise to sound natural
Strengthen logic
Revise title
Revise based on references
Swtich Formal/ Informal
Give example or cases
Write details
Pros and cons about subject
Thorough Writing
Simple edit
Give me ideas
Replace word
Discussions, questions about content
Write so the topic is clear
Write in required format
Revise to sound professional
Make it rich
Increase volume
Fill in gaps
Write based on this idea
Keep writing
Write the whole paper
Write conclusion
Content Drafting & Filling In
Language & Style Editing
Argument Development & Depth
Idea Generation & Pre-Writing
Structural & Format Refinement

Figure 3. Types of ChatGPT Prompts in Thorough Writing Conditions.

Figure 2 shows that, in the Quick Writing condition, students concentrated their ChatGPT usage on a few prompt types that directly expedite writing. The thickest arrows in the Quick Writing diagram, "Increase volume", represent a prompt that delegates substantial writing to ChatGPT. Another common prompt in Quick Writing was "Strengthen logic." While these two were among the most frequently prompted requests also in the Thorough condition, the low number of edits combined with their frequent use in the Quick condition suggests that students often offloaded the majority of the writing to the AI with minimal interaction when they are just trying to finish the paper as quickly as possible.

By contrast, the Quick Writing diagram shows very thin or missing arrows for prompts for deeper content development or iterative refinement. For instance, students in a hurry rarely asked ChatGPT to "Provide evidence or references," or "Give feedback." This infrequent use indicates that students did not engage ChatGPT for argumentative depth of their writing in the quick context. We also see limited use of interactive brainstorming or tutoring prompts. For example, "Give me ideas" or asking follow-up content questions were little asked in the Quick condition. These results indicate that students were less inclined to hold extended dialogues with the AI about the topic.

In contrast, the map for the Thorough Writing condition shows a wider variety of prompt types, as students worked to produce better quality writing. Although the most frequent prompt was "Increase volume," as in the Quick condition, students' requests in the Thorough condition ranged from initial idea generation (e.g., "Give me ideas," "Discussions about content") to advanced argument refinement ("Revise based on references," "Give examples or cases"). Compared to the Quick condition, Language and Style Editing was heavily used. Students actively asked ChatGPT to write details, revise title, and make simple edits here and there. They also sought to escape being caught for using ChatGPT (e.g., "Write to avoid plagiarism", "Revise to sound natural", "Write like a college student").

In total, students used 18 different types of prompts in the Quick condition whereas the same students used 30 different types of prompts in the Thorough condition. This indicates that they employed a far wider range of AI-related tasks when they were trying to produce a higher-quality product. Under Quick conditions, students tended to stick to a smaller set of prompt types often limited to generating text with minimal iteration. In contrast, the Thorough condition saw a broader variety of prompt use to support more comprehensive writing. For educators, it is worth noting that although students focused on quality interacted with the AI more extensively and in more complex ways, the most common use of ChatGPT across both conditions remained generating additional text to increase volume with minimal effort.

4. Discussion

From these results, we can infer that students made significantly more revisions to their drafts in the Thorough condition than in the Quick condition. However, this increased level of editing did not result in longer essays, as final word counts were statistically equivalent across the two conditions. Instead, it seems the extra effort invested during the Thorough condition led to improved content quality. Essays produced under the Thorough instructions showed greater depth and stronger logical coherence than those written under Quick instructions. These findings show that when students wrote for higher quality (Thorough), they engaged in more extensive revision and achieved more substantive content, even though the final text length was similar to the Quick outputs. These findings are consistent with process-oriented theories of writing [16,17,19], which focus on the importance of revision and reflection in the writing process. These theories view writing not as a linear but as a recursive process involving planning, drafting, and revising to develop and refine ideas. Our results align with this view, as the Thorough condition encouraged more extensive engagement with the text, leading to improved content quality. These findings were further supported by the subsequent research question.

Results for the third research question, which examined the type of prompts students used under each condition, showed noticeable differences. In the Quick condition, participants relied on a narrow range of prompt types (18 in total), whereas in the Thorough condition, the same students used a much broader repertoire (30 types). The most frequent prompt in both conditions was a request to "Increase volume," meaning the users asked ChatGPT to simply generate more text. This suggests that regardless of time or quality focus, students commonly used ChatGPT to expand their essays by lengthening them. However, under the Thorough condition, students employed a more diverse and complex set of prompts. When attempting to write a higher-quality paper, students not only asked for additional text but also leveraged ChatGPT for brainstorming ideas, refining arguments, and polishing language and style. By comparison, Quick condition students tended to stick to basic prompts that expedited completion and used ChatGPT with minimal iteration.

The findings carry significant implications for the development of future AI detectors and for writing practices in higher education. This study was motivated by the idea that focusing solely on the final product may be inadequate to accurately distinguish AI-generated text from human-authored work [6,12]. The results confirm that students in the Quick condition often produced text that superficially resembled that of their Thorough counterparts; however, their writing processes—including the types of prompts they used, the scope of their revisions, and their depth of engagement with the content—differed considerably. This discrepancy shows a key limitation of current AI detectors, which typically restrict their analysis to the final written product.

The findings from this study resonate with Salman and Alexandron's [21] study, which shows that students' behavioral patterns in digital learning environments can be uniquely identified through their "digital fingerprints" such as activity logs. Their work supports the idea that tracking revision behavior and prompt usage can provide meaningful evidence of engagement in students' writing assessments. By integrating process-based indicators, detectors could capture the micro-level decisions and revisions that reveal genuine cognitive investment. For example, tracking whether a writer made consistent revisions, requested deeper explanations, or added original commentary might indicate human effort. In contrast, a pattern of one-off, large-block text requests with minimal subsequent edits might suggest heavy AI reliance. Such data could come from revision histories, version control logs, or ChatGPT prompt records, enabling a more nuanced evaluation than final text analysis alone. Process-based AI detection can be considered a new paradigm in AI detection systems. Rather than relying exclusively on textual signatures or statistical patterns in a static document, next-generation detectors or word processors could incorporate process tracking. This approach would encourage more transparent and authentic writing practices, discourage "one-click" AI usage, and reduce the likelihood of false positives.

5. Recommendations and Limitations

In higher education, there is a growing call for a shift in how writing is taught. The challenges of detecting AI-generated text and the need for thoughtful approaches that balance academic integrity with innovation are ongoing, real-world issues that every educator faces [15, 23]. It is increasingly necessary to acknowledge that students will use ChatGPT for writing assignments, and even those who intend to use it only for assistive purposes may be tempted to generate text to increase word count or to avoid the strenuous process of completing an assignment quickly. This suggests that educators should guide students to use AI ethically, such as to support idea development, critical thinking, and revision [17,22], so students engage with AI in ways that enhance their own thinking and writing skills. By framing ChatGPT as a tool for brainstorming, elaboration, and refinement [24,25], educators can leverage its benefits while mitigating its use as a scaffold [26], wherein the student remains an active, reflective author rather than a passive consumer of AI-generated text [27].

The limitations of this study are as follows. First, the study involved 15 participants, which limits the generalizability of the findings. A small sample size may lead to results that are influenced disproportionately by individual differences or outliers and future research should replicate the study with a larger and more diverse sample to broaden the applicability of these conclusions. Second, the study did not investigate how students approach writing tasks without access to AI support, which makes it difficult to isolate the impact of ChatGPT specifically. Without a baseline for comparison, we cannot definitively determine the degree to which differences observed are directly attributable to the AI assistance. Future studies should include conditions in which students write entirely on their own to better clarify ChatGPT's influence.

6. Conclusion

Despite the limitations, the findings of this research reinforce a core principle in writing pedagogy that how students arrive at the final product is pedagogically telling. Educators and institutions may thus consider placing greater emphasis on students' editorial footprints [17], which may include the trail of planning, drafting, and revising actions a student takes. One practical implication is to incorporate process-based evaluation criteria. For instance, instructors could require students to submit evidence of their revision history or a summary of how they used AI during the writing process. By reviewing a document's revision history or the types of prompts a student used with ChatGPT, teachers can gain insight into the student's level of engagement and development. This approach aligns with long-standing principles of process-based assessment in writing education [28].

Author Contributions

Conceptualization, J.K. and Y.L.; methodology, J.K.; software, J.K.; validation, Y.L.; formal analysis, J.K.; investigation, J.K.; resources, Y.L.; data curation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, Y.L.; visualization, J.K.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding

Data collection and preliminary analysis were sponsored by Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A03062440).

Institutional Review Board Statement

Ethical review and approval were waived for this study due to not collecting personal data.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

Data is unavailable due to privacy restrictions.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Chaka, C., 2024. Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. Journal of Applied Learning and Teaching. 7(1), 115–126. DOI: https://doi.org/10.37074/jalt.2024.7.1.14

[2] Klein, A.B., 2023. New York City blocks ChatGPT at schools. Should other districts follow? Available from: https://www.edweek.org/technology/new-york-city-blocks-chatgpt-at-schools-should-other-districts-follow/2023/01 (cited 8 April 2025).

[3] Pearlman, J.B., 2024. Australian states block ChatGPT in schools even as critics say ban is futile. Available from: https://www.straitstimes.com/asia/australianz/australian-states-block–chatgpt-in-schools–even-as-critics-say-ban-is-futile (cited 10 October 2024).

[4] Yu, H., 2023. Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. Frontiers in Psychology. 14, 1–12. DOI: https://doi.org/10.3389/fpsyg.2023.1181712

[5] Stone, B.W., 2024. Generative AI in higher education: Uncertain students, ambiguous use cases, and mercenary perspectives. Teaching of Psychology. Advance online publication. DOI: https://doi.org/10.1177/00986283241305398

[6] Chaka, C., 2023. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching. 6(2), 94–104. DOI: https://doi.org/10.37074/jalt.2023.6.2.12

[7] Desaire, H.B., Chua, A.E., Kim, M.G., et al., 2023. Accurately detecting AI text when ChatGPT is told to write like a chemist. Cell Reports Physical Science. 4(11), 101672. DOI: https://doi.org/10.1016/j.xcrp.2023.101672

[8] Gao, C.A., Howard, F.M., Markov, N.S., et al., 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digital Medicine. 6(1), 75. DOI: https://doi.org/10.1038/s41746-023-00819-6

[9] Tian, E.B., 2023. GPTZero: Perplexity, burstiness, and statistical AI detection. Available from: https://gptzero.me/news/perplexity-and-burstiness-what-is-it/ (cited 1 March 2025).

[10] Turnitin, 2023. To ban or not to ban AI writing in schools? Available from: https://www.turnitin.com/blog/to-ban-or-not-to-ban-ai-writing-in-schools (cited 1 March 2024).

[11] Dwyer, M., Laird, E.B., 2024. Up in the air: Educators juggling the potential of generative AI with detection, discipline, and disrupt. Center for Democracy and Technology. Available from: https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf (cited 10 March 2025).

[12] Bender, E.M., Koller, A.B., 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5–10 July 2020. pp. 5185–5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463

[13] Goldstein, S.B., Kern, R.B., 2023. Evaluating AI-generated writing: Challenges and emerging solutions. Journal of Educational Technology. 46(2), 175–190. DOI: https://doi.org/10.1016/j.asw.2024.100899

[14] Weber-Wulff, D.B., Anohina-Naumeca, A.B., Bjelobaba, S., et al., 2023. Testing of detection tools for AI-generated text. International Journal for Educational Integrity. 19. DOI: https://doi.org/10.1007/s40979-023-00146-z

[15] Lancaster, T., 2023. Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution? International Journal for Educational Integrity. 19(1), 10. DOI: https://doi.org/10.1007/s40979-023-00131-6

[16] Emig, J., 1977. Writing as a mode of learning. College Composition and Communication. 28(2), 122–128. DOI: https://doi.org/10.2307/356095

[17] Flower, L.B., Hayes, J.R., 1981. A cognitive process theory of writing. College Composition and Communication. 32(4), 365–387. DOI: https://doi.org/10.2307/356600

[18] Bereiter, C.B., Scardamalia, M.B., 1987. The Psychology of Written Composition. Routledge: New York, NY, USA. DOI: https://doi.org/10.4324/9780203812310

[19] Sommers, N., 1980. Revision strategies of student writers and experienced adult writers. College Composition and Communication. 31(4), 378–388. DOI: https://doi.org/10.2307/356588

[20] Hayes, J.R., 2012. Modeling and remodeling writing. Written Communication. 29(3), 369–388.

[21] Salman, A.B., Alexandron, G.B., 2024. The digital fingerprint of learner behavior: Empirical evidence for individuality in learning using deep learn-ing.

Computers and Education: Artificial Intelligence. 7, 100322. DOI: https://doi.org/10.1016/j.caeai.2024.100322

[22] Myhill, D.B., Jones, S.B., 2007. More than just error correction. English Teaching: Practice and Critique. 6(2), 8–31. DOI: https://doi.org/10.1177/0741088312451260

[23] Waltzer, T.B., Pilegard, C.B., Heyman, G.D., 2024. Can you spot the bot? Identifying AI-generated writing in college essays. International Journal of Educational Integrity. 20. DOI: https://doi.org/10.1007/s40979-024-00158-3

[24] Dilekli, Y.B., Boyraz, S.B., 2024. From "Can AI think?" to "Can AI help thinking deeper?": International Journal of Modern Education Studies. 8(1), 49–71. DOI: https://doi.org/10.51383/ijonmes.2024.316

[25] Songkram, N., Chootongchai, S., Keereerat, C., et al., 2024. Potential of ChatGPT in academic research: Exploring innovative thinking skills. Interactive Learning Environments. 1–23. DOI: https://doi.org/10.1080/10494820.2024.2375342

[26] Vygotsky, L.S.,1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press: Cambridge, MA, USA. DOI: https://doi.org/10.2307/j.ctvjf9vz4

[27] Grinschgl, S., Neubauer, A.C., 2022. Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future. Frontiers in Artificial Intelligence, 5, pp.1–6. DOI: https://doi.org/10.3389/frai.2022.908261

[28] White, E.M., 1994. Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance (Revised and expanded ed.). Jossey-Bass Publishers: San Francisco, CA, USA.