Designing and Evaluating AI Margin Notes in Document Reader Software

Nikhita Joshi*
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
nvjoshi@uwaterloo.ca

Daniel Vogel
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
dvogel@uwaterloo.ca

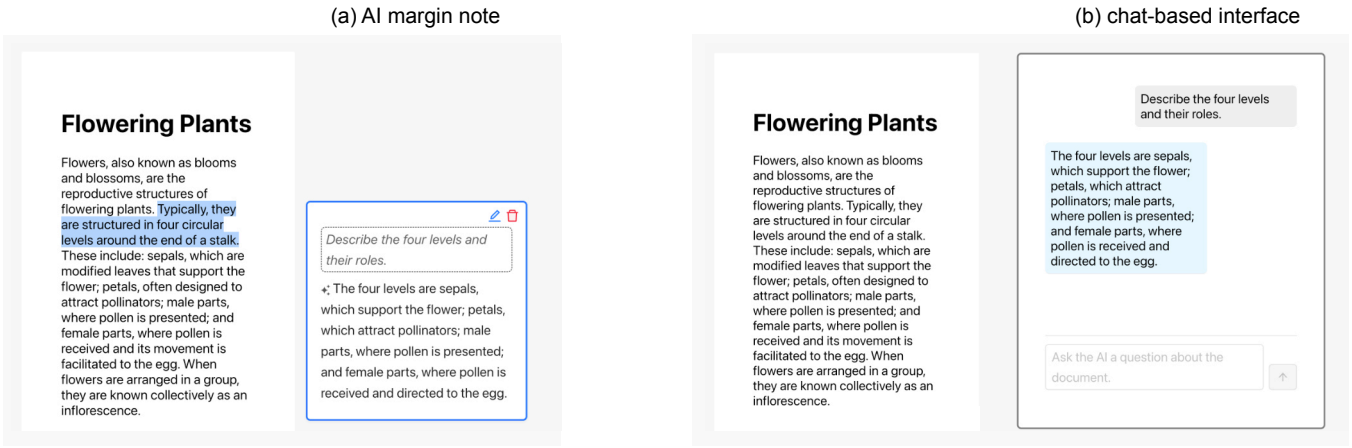

Figure 1: Different ways to leverage LLM capabilities in document reader software: (a) we propose AI margin notes that are integrated with the document text unlike (b) chat-based interfaces that are separated from the document text.

Abstract

AI capabilities for document reader software are usually presented in separate chat interfaces. We explore integrating AI into document comments, a concept we formalize as AI margin notes. Three design parameters characterize this approach: margin notes are integrated with the text while chat interfaces are not; selecting text for a margin note can be automated through AI or manual; and the generation of a margin note can involve AI to various degrees. Two experiments investigate integration and selection automation, with results showing participants prefer integrated AI margin notes and manual selection. A third experiment explores human and AI involvement through six alternative techniques. Techniques with less AI involvement resulted in more psychological ownership, but faster and less effortful designs were generally preferred. Surprisingly, the degree of AI involvement had no measurable effect on reading comprehension. Our work shows that AI margin notes are desirable and contributes implications for their design.

*Also with LISN, Université Paris-Saclay, CNRS, Inria.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790786

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Interaction techniques**.

Keywords

interaction techniques, controlled experiments, large language models, generative AI, note-taking

ACM Reference Format:
Nikhita Joshi and Daniel Vogel. 2026. Designing and Evaluating AI Margin Notes in Document Reader Software. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3790786

1 Introduction

Taking notes while reading documents is a common active reading strategy that can be done in two primary ways. First, people can write notes that are decoupled from the document text by writing on a separate piece of paper. Second, people can write notes that are integrated with the document text [1], for example, by writing in the margins (*marginalia*). Such 'margin notes' have been common practice for centuries, as they provide a space for readers to summarize, paraphrase, explain unfamiliar concepts, make connections to existing knowledge, and even express personal opinions [31, 53]. This can be especially beneficial when reading for educational purposes [8], as it provides space for people to work through difficult sections and quickly re-access their thoughts afterwards within a shared context [1, 42]. With digital documents, margin notes can

be added as digital ink with a pen (e.g., [54]), but a more common way is leaving *comments* when reading documents in systems like Google Docs, Microsoft Word, and Adobe Acrobat Reader [63]. This is usually done by selecting text within a document and writing in a text box positioned beside it.

Large language models (LLMs) can support reading documents to summarize, reword, and better understand unfamiliar concepts [26, 27, 39]. The generated output serves a purpose like margin notes, yet the capabilities of LLMs have not been integrated into commenting features of document reader software. Instead, most LLM-enhanced document readers use a separate, chat-like interface disconnected from the document. There is an opportunity to improve how LLMs are used in this context by integrating their capabilities directly into 'margin note' comments that are linked to the document text.

We explore the design of "AI margin notes" that leverage LLMs to enhance comments in document reader software through three controlled experiments. Every experiment required participants to read short, non-fiction documents while interacting with an LLM, primarily within comments, and answer reading comprehension questions two hours later. Each experiment focused on a different design parameter. First, we compared AI margin notes to traditional, chat-based prompting to better understand the effects of *integration*. Second, we compared manually selecting text that an AI margin note is associated with, to automatically selecting text through an LLM-powered assistant to better understand the effects of *selection automation*. Third, we explored different AI margin note techniques that leverage LLMs in different ways to understand the effects of *human and AI involvement*. Specifically, AI margin notes that generate summaries, fill-in-the-blank exercises, responses to specific prompts, practice short answer questions, and feedback on how written text can be improved.

Our results indicated that AI margin notes were preferred over chat interfaces and that selecting text for AI margin notes should be done manually. AI margin note techniques with varying levels of human and AI involvement were valued for different reasons. For example, techniques with more human involvement were typically associated with more psychological ownership, but techniques with more AI involvement were faster, less effortful, and generally preferred. Our work contributes:

- the idea of AI margin notes: comments that are enhanced with LLMs to support note-taking that is integrated with the document text; and
- empirical results from three experiments, each focusing on a specific design parameter, demonstrating that AI margin notes are a desirable feature for document reader software.

2 Background and Related Work

AI margin notes relate to existing literature in psychology about the benefits and challenges of note-taking while reading, and other techniques that have used LLMs to improve reading and note-taking. We focus specifically on relevant work focused on reading, rather than taking notes while attending a lecture or watching a video.

2.1 Note-Taking while Reading

There are many types of notes that readers can create, margin notes being one of them. Research in psychology suggests that the general activity of note-taking is beneficial for two main reasons [37]. First, notes act as external storage for information, which can be re-read to reinforce memory (the *storage function*). Second, the act of creating notes can facilitate learning as it requires readers to pay more attention to the material to process and organize ideas (the *encoding function*).

The encoding function is especially beneficial when the reader processes the text at a deeper level [17, 36], for example, by connecting it to prior knowledge and experiences. However, many people opt for shallower and less effective note-taking strategies. For example, Bretzing and Kulhavy's analysis of students' note-taking activities [11] revealed that students tend to take verbatim notes that repeats text from documents. In another study [10], they showed that students who took verbatim notes performed worse on a test than those who used deeper note-taking strategies by writing their own summaries or by paraphrasing text from the document.

Note-taking is a complex activity that requires significant cognitive effort as readers must coordinate and frequently switch between reading and writing activities. Writing takes more time than reading, and excessive delays between reading activities as a result of note-taking can hinder comprehension. Therefore, readers often experience significant mental and temporal demand, even when they are reading and taking notes without any time limits [52]. This can become especially tiring with longer [38] or poorly-formatted documents [49]. Although researchers have argued that increased cognitive effort can benefit learning [5–7] and memory [59], it does not always lead to improved learning outcomes, for example, if the task is too frustrating or if the learner lacks motivation [25].

2.2 Reading and Note-Taking with LLMs

Recent work has investigated how LLMs can improve user experience and comprehension while reading and taking notes. For example, text simplification techniques that turn complex documents into simplified versions [3, 27] and make documents easier to skim [26] can improve reading comprehension and reduce workload. Users of systems like ChatGPT, Adobe Acrobat [2], Google Notebook LM [23], NoteGPT [48], and ChatPDF [15] can ask questions, summarize, and write notes about documents. However, few investigations examine the effect on factors like reading comprehension, reading duration, workload, and preferences.

Kreijkes et al. [39] asked high school students to try different note-taking techniques to study for reading comprehension tests that took place three days later: note-taking independently and note-taking while having access to an LLM-powered chatbot to ask questions. Both of these note-taking techniques were compared to a baseline of asking the chatbot questions about the document, without any note-taking. Their results showed that both note-taking techniques led to better performance than just asking the chatbot questions about the document, suggesting that using LLMs in a more cognitively engaging way (i.e., with some note-taking) can improve reading comprehension. However, they did not compare the two note-taking conditions, making the effect of LLM use on note-taking unclear.

Although Kreijkes et al. found that just asking the LLM questions led to poorer test performance, it was preferred by participants as it was perceived to be more enjoyable and less effortful. Such findings

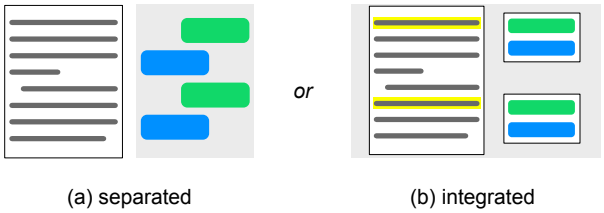

Figure 2: Integration: (a) a chat-based interface is not integrated since it is separated from the associated document text and (b) AI margin notes are integrated since they are associated to specific document text. *In this and the following two figures: green denotes human-written text, like a prompt, and blue represents AI-generated text, like a response. Yellow indicates text that the AI margin note is linked to.*

have been reproduced in other studies focused on the effects of LLMs on learning more broadly. Systems like ChatGPT can allow for personalized learning experiences, which can improve academic performance, reduce mental effort, and motivate learners [20, 61], however, learners may over-rely on these systems [61]. A balance may be to encourage more cognitive engagement. For example, ChatPRCS [62] generates practice reading comprehension questions for students, which increased mental load but improved reading comprehension. Similarly, CoAsker [41] encourages students to generate practice questions with an LLM-powered assistant, which were displayed in a side panel much like notes beside the document. When compared to generating practice questions without assistance, receiving questions that were generated by the assistant led to higher reading comprehension scores.

To our knowledge, no prior work has thoroughly explored integrating LLMs into the commenting feature of document reader software. The closest prior work are prototype mockups in Melin-Higgins' bachelor's thesis [45], in which users select document text to issue pre-determined prompts to a hypothetical LLM-powered assistant. Responses appear as 'sticky notes' beside the selected text, like margin notes, or they appear underneath the selected text by modifying the document structure. A very small 3-person study showed a preference for the margin note style. The study also only focused on usability and user preferences, and did not consider factors like reading comprehension. Furthermore, the prototypes were partially functioning Figma mockups, and none explored different levels of human and AI involvement.

Based on this research, AI margin notes could hinder or help readers. Creating them may limit deeper processing of the document text and lower comprehension. However, some AI margin note creation techniques may make note-taking less mentally demanding, so readers focus more on the underlying text while reading. Some techniques may even help readers take non-verbatim notes while improving motivation and cognitive engagement. Therefore, it is important to understand the effect of AI margin notes, and note-taking with LLMs more generally, on factors like test performance, workload, and user preferences. Our work contributes these important insights, which have been lacking in existing literature.

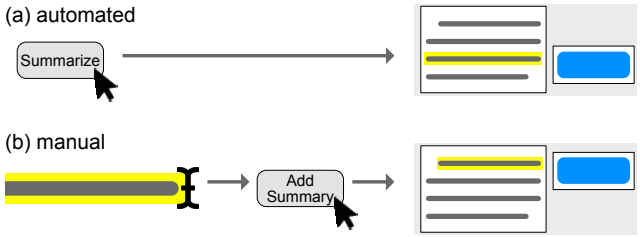

Figure 3: Selection automation: (a) text can be automatically selected and multiple AI margin notes can be created at once and (b) the user can manually select text to create an AI margin note.

3 AI Margin Notes

We focus on three design parameters of AI margin notes: *integration*, *selection automation*, and the level of *human and AI involvement*.

3.1 Integration

The primary difference between AI margin notes and chat-based tools is integration. Specifically, AI margin notes enhance *comments* in document readers. Comments are anchored to specific text, making them integrated into the document, which contrasts with chat-based tools that are placed in a separate side panel with the content disconnected from the context (Figure 2). A separate interface for interacting with an AI assistant may suffice for general questions about the document or to receive overall summaries. However, readers also direct the AI assistant to specific parts of the text to contextualize their prompt [39]. Separating these responses from the text can be inefficient.

First, *referring to specific parts of a document while prompting requires additional work* [44]. Consider prompting an LLM to simplify a paragraph in a scientific document (e.g., [3, 27]). With a chat-based interface, the user must carefully formulate a prompt to refer to the specific paragraph (e.g., "simplify the third paragraph in section 3"). Or, the user must select, copy, and paste the paragraph into the chat and write a prompt that uses *deictic words* to refer to the paragraph instead (e.g., "simplify *this*"). With an AI margin note, the interaction is simpler as user can use deictic words without a separate copy and paste stage: they just select the paragraph and type "simplify this" where it appears in the document.

Second, *shifting attention to a separate interface may distract from reading*. Note-taking requires frequent switching between reading and writing [2, 52], and researchers suggest that note-taking should "interrupt reading as little as possible" [42]. Having readers frequently switch between separate reading and writing interfaces may further increase cognitive load [29]. These switching costs could be mitigated by presenting the prompting interface alongside the specific text the user is reading.

Third, *referring back to specific responses requires additional work*. For example, suppose a user prompted a chat-based LLM interface to get an explanation for an unfamiliar concept in a document. The explanation may be remembered in the short term, but if the user revisits the document much later, they must scroll through a lengthy and poorly organized chat history to find it, which may be especially difficult for readers with a lower working memory capacity to do [55]. Linking responses directly to the relevant text avoids this issue.

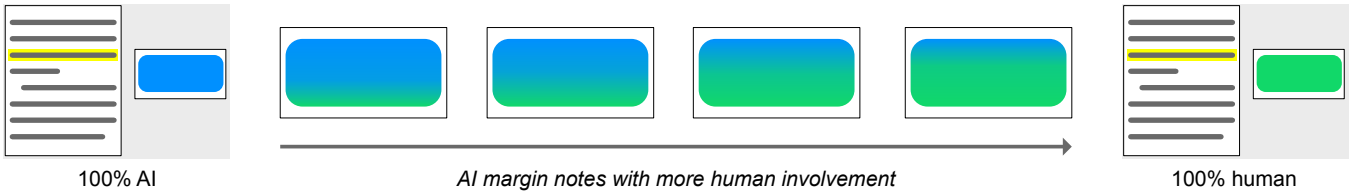

Figure 4: Levels of human and AI involvement: traditional margin notes consist of text that is 100% written by a human, but an AI margin note could consist of text that is 100% generated by AI, or could involve more human-generated text.

Some systems, like Adobe Acrobat Reader, support clicking on chat responses to highlight relevant parts of the document, however, this still requires scrolling through chat history. Alternatively, the user could re-issue the prompt, but this wastes time and resources. Of course, users could copy responses they wish to save and paste them in comments [39], but this requires additional copy and paste steps that would not be needed with an AI margin note.

3.2 Selection Automation

AI margin notes are associated with specific text, which defines the context for the prompt and a location to display the note. Specifying text for the note could be done automatically by the AI assistant or manually by the user (Figure 3). For example, consider summarization, a common task readers do in conventional margin notes [31, 53] and with LLM systems [39, 45]. When using LLMs in document reader software, readers can generate a summary of the entire document automatically by pressing a single button [2], or they can manually specify which parts of the document need to be summarized by copying and pasting text into the chat. When the reader just presses a button, the LLM decides which information is important and worth including in the summary, but when the reader copies text to summarize, they are deciding which parts of the document are most important. Prior work on text highlighting suggests that this decision process could improve learning outcomes and reading comprehension, but it could also increase mental effort [33, 66].

Automating the selection of text to create AI margin notes may also impact psychological ownership, feelings of the learning experiences *belonging* to them [50]. Prior work shows that fostering psychological ownership, which can be achieved by giving learners more *control* over their learning [50, 51], can encourage more active learning [24] and motivates learning [56]. Automatically generating AI margin notes reduces effort, but potentially at the risk of lowering comprehension and psychological ownership.

3.3 Human and AI Involvement

Currently, commenting in document reader software is intended for text that is entirely written by a human. Specifically, the user must formulate their own ideas and type into a text box. At the other extreme, pressing a "Summarize" button in a document reader [2] produces text that is entirely written by the LLM without any guidance from the human. However, there are techniques that require involvement from both the human and the LLM. For example, a chat-based interface requires the user to provide a specific instruction to the LLM, such as "summarize the text for someone in high school." Here, the user forms goals and sub-tasks [57] to exert some control over the output [35], and the LLM produces it. These roles can be reversed too, for example, the LLM could 'prompt' the user to respond to practice reading comprehension questions [41, 62].

AI margin notes can also vary in how much human and AI involvement they require to produce the final comment text (Figure 4). Techniques that require more human involvement likely require more cognitive engagement than those with more AI involvement. This may be beneficial for reading comprehension [5–7, 59], may discourage readers from taking verbatim notes [10], and may even improve feelings of psychological ownership [34, 35]. However, too much cognitive engagement can frustrate and discourage learners [25], which may be mitigated by increasing AI involvement.

4 Experimental Method

We conducted three experiments to explore these three design parameters of AI margin notes. Specifically, we were interested in understanding how integration, selection automation, and human and AI involvement affect reading comprehension, duration, psychological ownership, task workload, and user preferences. All experiments used the same experimental method and were conducted using the Prolific crowdsourcing platform,[1] but with different participants for each experiment. Note our protocol was reviewed and approved by our institution's Research Ethics Board.

4.1 Task

The primary experimental task was composed of two stages. First, a *reading stage*, where participants read non-fiction documents that were approximately 500 words each. For each document, they interacted with an LLM assistant using a different technique that represents a variation within the design parameter under evaluation. Second, a *test stage*, where participants completed reading comprehension tests about each document. Each test consisted of six multiple choice questions. The documents and questions were developed by Wallace et al. [60] and were designed by a learning and reading specialist to be suitable for an eighth grade reading level.[2] This is representative of documents targeted for the general public [47].

4.2 Apparatus

Our experimental system was a custom Node.js and React web application that implemented two types of interfaces, one for each

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

[2] Note that we slightly modified the wording of some questions, as we noticed that some questions could be successfully answered by looking at the wording of other questions. Our versions of these questions and the associated documents are included in the supplementary materials. At the time of writing this paper, Wallace et al.'s license agreement permits reuse, modification, public display, and redistribution for non-commercial research purposes.

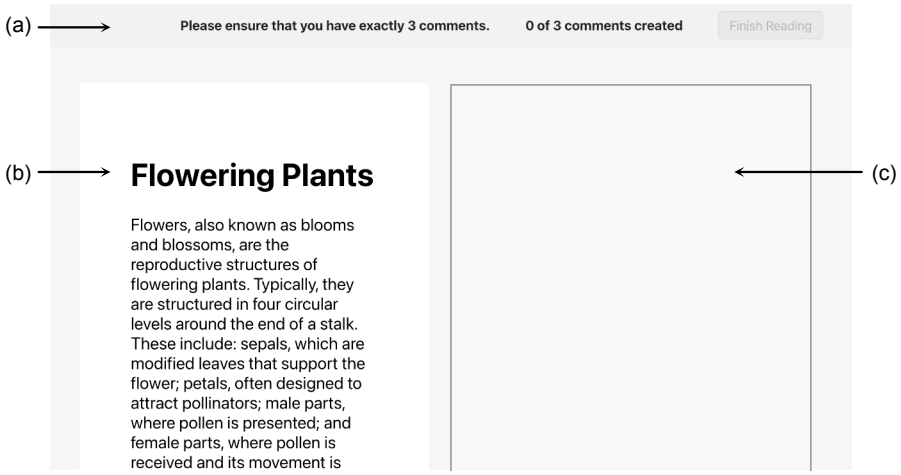

Figure 5: Experimental reading interface: (a) toolbar containing instructions, (b) document to read, and (c) space for specific design variations to be displayed.

stage: the reading interface, where participants read a document and interacted with an LLM using a specific technique (Figure 5); and the testing interface, where participants answered time-bounded, multiple-choice reading comprehension tests. *The supplementary video demonstrates both interfaces.*

4.2.1 Reading Interface. The document was displayed on the left side of the screen. The specific design variation of the experimental condition was displayed beside it (details provided in each individual experiment). All design variations that involved an LLM used GPT 4.1 mini.[3] The top of the screen contained a toolbar indicating how many comments or responses they still had to complete and a blue "Finish Reading" button to end the trial.

4.2.2 Testing Interface. The testing interface displayed six multiple choice questions at the centre of the screen. At the top of the screen was a toolbar that displayed how many questions the participant had answered, a countdown timer (displayed numerically and as a progress bar that shrank every second), and a blue "Finish Test" button that could be pressed to end the test.

4.3 Procedure

Participants received a link to an experimental web application through Prolific. They had to use a desktop or laptop computer, which was strictly enforced through the web application. First, participants read a consent form, which detailed inclusion and exclusion criteria, data handling procedures (anonymized and stored on encrypted hard drives), and remuneration. After providing informed consent, participants entered basic demographic information and read instructions.[4] The instructions described the general nature of the reading and testing stages, and details on how to use the specific techniques being tested.

Next, they completed the reading stage. There was no time limit, but to focus on the effect of the different techniques rather than the number of responses, participants had to leave 3 AI margin notes (or receive 3 responses from the AI assistant, depending on the technique). After reading the document, they answered 10 questions about their experience and 1 question about their prior knowledge of the document's content. They repeated this for the other documents, then answered 2 questions about their overall preferences.

Two hours later, the participant was invited to return for the test stage. Given the lower reading difficulty of the documents, we increased the difficulty of the test by making it closed-book and restricted to 60 seconds. After completing all tests, participants described other study aids they used, such as taking a screenshot of the document or writing notes outside of the reading interface.

4.4 Design

All experiments use within-subjects design since the documents were relatively short, and we wanted participants to compare their experiences with each technique. There is one independent variable, CONDITION, which was randomly assigned. For a single CONDITION, one of six documents from Wallace et al. [60] was randomly-assigned. For increased internal validity, a single document was not restricted to a single condition and could be assigned to any condition.

4.5 Measures

Reading comprehension and duration were calculated from logs, and all other metrics were calculated from subjective questionnaires completed after the reading stages. With the exception of rankings, these used a 0-100 interval range.[5]

4.5.1 Reading Comprehension. This represents the number of questions that were correctly answered during the test stage (0-6 range). Analyzing *Reading Comprehension* gives insights into the effects of AI margin notes on overall learning, and whether techniques that encourage more human involvement improve learning outcomes.

[3]The system prompts used to generate responses are included in the supplementary materials.

[4]The demographics questionnaire and all instructions are included in the supplementary materials.

[5]The supplementary materials contains all questions.

Table 1: Experiment 1 demographics.

Gender		Age		Education	
Men	16	25-34	4	High School	2
Women	10	35-44	14	Some University (no credit)	5
		45-54	3	Bachelor's Degree	14
		55-64	3	Professional Degree Beyond Bachelor's	1
		65-74	2	Master's Degree	3
				Doctorate Degree	1

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	6	Daily	1	Daily	10	Daily	3
Weekly	11	Weekly	6	Weekly	9	Weekly	9
Monthly	7	Monthly	3	Monthly	5	Monthly	3
Less than Monthly	2	Less than Monthly	5	Less than Monthly	5	Less than Monthly	5
		Never	11	Never	1	Never	6

4.5.2 Duration. This is the time taken in minutes to complete the reading stage. Some AI margin notes that require more manual effort or human involvement likely take longer to complete.

4.5.3 Psychological Ownership. This was measured by asking two quantitative questions about *Personal Ownership* and *Responsibility* [13]. The average of the two create a composite measure, as done in prior work [34, 35]. The internal consistency reliability score, which describes how consistently different items on a questionnaire describe the same underlying concept, was high for all experiments ($.80 \leq \alpha \leq .90$), suggesting that the composite measure was appropriate to use for data analysis. As described previously, this is an important measure since fostering psychological ownership can improve learner experiences.

4.5.4 Task Workload. We used factors from the NASA-TLX [30]: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. Techniques that are more manual or that require more human involvement may involve more mental demand and effort, which could be beneficial for learning [5–7, 59], however, this may not happen if they are too frustrating [25].

4.5.5 Preferences. We asked about *Frequency of Use*, a question from the system usability scale (SUS) [12] representing how frequently the participant would use the feature if made available to them. After trying all techniques, participants also gave each CONDITION a *Ranking*, where 1 was the best, 2 was the second best, and so on. Ties were allowed. To establish an *Overall Ranking* of conditions, we use the Condorcet voting method [65]. An overall rank of 1 means that technique 'defeats' all others in pairwise comparisons. An overall rank of 2 means that technique 'defeats' all others except for the technique ranked first, and so on. This is important to study, as users would likely want to keep using techniques that they like, which is especially useful if it also improves their learning outcomes.

4.5.6 Other Metrics. We triangulate these measures with other data, such as characteristics of participant prompts and the length and location of selected text. These metrics are specific to each experiment, so we introduce them with their results.

5 Experiment 1: Integration

The goal of this experiment is to understand the effect of integration that is achieved with AI margin notes. Participants wrote prompts using a traditional, chat-based interface (CHAT), or an AI margin note (NOTE).

5.1 Participants

We recruited 29 participants through Prolific. Participants were restricted to the United States and Canada, and those who had completed 2,500 previous tasks on the platform with a 99-100% approval rating. Three participants (10%) were removed for using other study aids prior to the test, leaving 26 valid responses (Table 1). All self-reported proficiency in reading in English. Each participant received $10, plus a $5 bonus for scoring in the top 25% on the comprehension test, to encourage more conscientious active reading. The experiment took roughly 25 minutes in total.

5.2 Apparatus

The reading interface had two variations. For the chat-based prompting technique, a chat interface was displayed to the right of the document. Here, participants could type prompts in a text box and press a blue "Submit" button. Their prompts appeared like a chat message on the right in a grey bubble, and LLM responses were displayed on the left in a light blue bubble (Figure 6a).

To make more direct comparisons between the two variations, we had to create an AI margin note technique where participants type prompts and receive responses from an LLM. To create an AI margin note, the participant selected text in the document using their cursor, which caused a blue "Comment" button to appear. Clicking this created a text box to appear that was anchored to part of the document text and resembled a Google Docs comment (Figure 6b). The participant could type a prompt and press a blue "Confirm" button. After a few seconds, the LLM provided a response. The participant's prompt appeared at the top of the comment in grey, italicized text with a dashed border around it, and the response appeared below in black text, prepended with a black 'sparkle' icon. This styling clearly distinguishes human-written text (no icon) from AI-written text (with a 'sparkle' icon) and main comment text (black text) from additional context that helped generate the

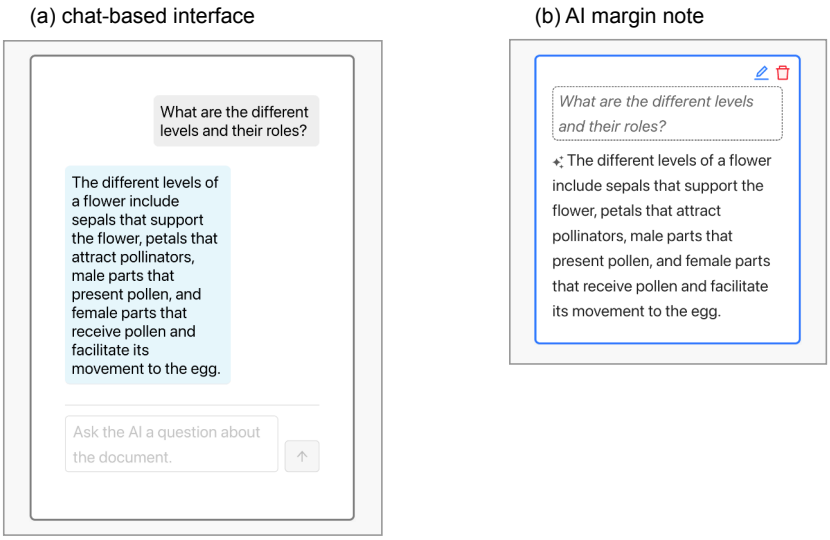

Figure 6: Techniques tested in Experiment 1: (a) a chat-based interface and (b) AI margin notes.

comment (grey, italicized text with a dashed border). Participants could edit their prompt to receive new output and they could delete their comments. Clicking on individual comments highlighted the corresponding text in the document.

5.3 Results

We use Wilcoxon signed-rank tests to investigate the effects of CONDITION on the various measures. To streamline the presentation of results, *details of these statistical tests are shown in Table A.1* and *all data is included in the supplementary materials.*

Overall, there were no significant effects of CONDITION on *Reading Comprehension*, *Duration*, *Psychological Ownership*, any of the workload-related factors, and *Frequency of Use*.

5.3.1 Preferences. Although there were no differences for the aforementioned measures, most participants preferred prompting with AI margin notes. From the Condorcet voting method, we observed that NOTE received an *Overall Ranking* of 1, and CHAT received an *Overall Ranking* of 2. Notably, the majority of participants (17, 65%) assigned NOTE a *Ranking* of 1, and CHAT a *Ranking* of 2 (16, 62%).

To better understand why participants preferred NOTE over CHAT, we examined participants' free-form responses. Several (12, 46%) mentioned how NOTE was easier and more intuitive than CHAT. They described how they *"liked [not] having to move to a separate chat"* (P21), how *"having the comment embedded right into the document [was] handy and easy to reference along with the actual text"* (P5), and how *"selecting text [was] more natural when one has a question related to that part of the text"* (P14).

5.3.2 Prompt Wording. These responses regarding the ease of referring to specific parts of the document were corroborated by examining how participants worded their prompts. Participants did not have to write prompts that were as specific when they used NOTE, perhaps due to the increased specificity enabled by text selections. Notably, 29 prompts (37%) written with NOTE referred to specific nouns in the document using deictic words (e.g., "this," "they," "it," and "here"). With CHAT, deictic words were less frequent (17, 22%) and most instances referred to specific nouns from previously-entered prompts (10, 13%). For example, consider P3 and P21, who both asked follow-up questions about materials used to make wagon wheels. P21, who read this document with CHAT, asked *"what were wagon wheels made of before iron?"* But P3, who used NOTE, instead selected text that discussed wagon wheel material, and asked *"how much of an impact did **this** have for the world?"*

5.4 Summary

Overall, our results suggested that the integration of AI margin notes is highly desirable and preferable over chat-based interfaces as it prevented switching between interfaces; it was easier to reference individual comments later; and it was easier to to ask questions using deictic words to refer to specific nouns described in the selected text. In this experiment, associating text to the AI margin note was done *manually* by the participant. However, selecting text could also be done *automatically* by the LLM, which we explore further in the following experiment.

6 Experiment 2: Selection Automation

After learning that the integration of AI margin notes through text selections was desirable, we then wanted to learn *how* such margin notes should be associated to specific text selections. The goal of this experiment is to understand the effect of selection automation when creating AI margin notes. Participants either manually selected what text should be summarized (MANUAL), or pressed a button to let the AI assistant place three summary AI margin notes in the document (AUTOMATIC).

6.1 Participants

Using the same inclusion criteria as Experiment 1, we recruited 32 new participants on Prolific. Two participants (6%) were excluded for not attempting to answer any reading comprehension questions, or for using other study aids, leaving 30 valid responses (Table 2).

Table 2: Experiment 2 demographics.

Gender		Age		Education	
Men	18	25-34	4	High School	4
Women	12	35-44	15	Some University (no credit)	7
		45-54	8	Bachelor's Degree	9
		55-64	1	Master's Degree	10
		65-74	1		
		75+	1		

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	4	Daily	1	Daily	8	Daily	2
Weekly	10	Weekly	3	Weekly	12	Weekly	8
Monthly	7	Monthly	4	Monthly	5	Monthly	4
Less than Monthly	7	Less than Monthly	3	Less than Monthly	5	Less than Monthly	11
Never	2	Never	18	Never	2	Never	5

All self-reported English reading proficiency. As in Experiment 1, participants received $10 with a $5 bonus incentive for top-25% comprehension test performance. The experiment took approximately 25 minutes in total.

6.2 Apparatus

Both variations focused specifically on producing AI margin notes that summarized text as this is a representative task that is currently done manually, by the user copying and pasting specific parts of a document into a chat-based interface, or automatically, by pressing a button. For the automatic selection technique, the top toolbar displayed a blue "Generate Comments" button in the top left corner (Figure 7a). Pressing this caused three AI margin note comments that were linked to the document text to appear to the right of the document. Each summarized specific selections from the document, and was displayed in black text prepended with a black 'sparkle' icon. As before, participants could click individual comments to highlight the corresponding text, but they could not delete or edit individual comments. Instead, they could press the "Generate Comment" button again to regenerate all comments.

The manual selection technique worked like the previous experiment (Figure 7b): the participant selected text from the document and pressed a blue "Comment" button to generate a summary comment (appearing after a few seconds). The styling of the comment was the same, except that participants could also delete or regenerate individual comments. The participant had to manually create three summary comments.

6.3 Results

As before, we use Wilcoxon signed-rank tests where applicable. *Details of statistical tests are shown in Table A.2.*

We did not observe any significant effect of CONDITION on *Reading Comprehension*, so we focus our results on other metrics.

6.3.1 Duration. Participants were 39 seconds slower when they manually selected text (Figure 8a). There was a significant effect of CONDITION on *Duration*, revealing that MANUAL (MDN = 3.43, IQR = 2.15) was slower than AUTOMATIC (MDN = 2.78, IQR = 2.14).

6.3.2 Psychological Ownership. Though slower, participants generally felt more psychological ownership when they manually selected text (Figure 8b). A significant effect of CONDITION on *Psychological Ownership* revealed that participants felt more *Psychological Ownership* with MANUAL (MDN = 50.25, IQR = 46.62) than AUTOMATIC (MDN = 6.75, IQR = 10.88). This was supported by free-form responses like *"[manually selecting text] requires me to focus more on the task, [which] gives me more responsibility"* (P20).

6.3.3 Task Workload. Participants felt like they performed better when selecting text manually, despite it requiring more effort. Specifically, there was a significant effect of CONDITION on *Performance* (Figure 8c), suggesting that participants felt like they were better at creating AI margin notes with MANUAL (MDN = 88, IQR = 24.25) than AUTOMATIC (MDN = 80.5, IQR = 49.5).

A significant effect of CONDITION on *Effort* (Figure 8d) suggested that participants felt more *Effort* for MANUAL (MDN = 35, IQR = 41.25) than AUTOMATIC (MDN = 14, IQR = 38.75). This was supported by free-form responses like *"in being able to highlight on my own, I also feel as if I did some of the work and not just nothing"* (P25). We did not observe any differences between MANUAL and AUTOMATIC for *Mental Demand*, *Physical Demand*, *Temporal Demand*, and *Frustration.*

6.3.4 Preferences. Participants generally seemed to prefer generating AI margin notes manually. There was a significant effect of CONDITION on *Frequency of Use* (Figure 8e), suggesting a preference to use MANUAL (MDN = 79.5, IQR = 47.5) more frequently than AUTOMATIC (MDN = 55, IQR = 68.75). For *Overall Ranking*, we observed that participants tended to prefer MANUAL, which was ranked first. Specifically, 23 participants (77%) gave MANUAL a *Ranking* of 1, and 21 (70%) gave AUTOMATIC a *Ranking* of 2.

A few (4, 13%) mentioned how MANUAL provided more incentive to read the document, for example: *"generating all summaries at once will encourage users not to actually read the document for themselves. Generating individual summaries after I selected text meant I [had] to actually read the document"* (P10). The majority (16, 53%) appreciated MANUAL for the increased control it enabled, for example: *"I felt a little more control [when] I was able to pick what I wanted to be summarized. [When] I just pressed one button, it summarized everything, but not necessarily the areas I wanted it to"* (P22).

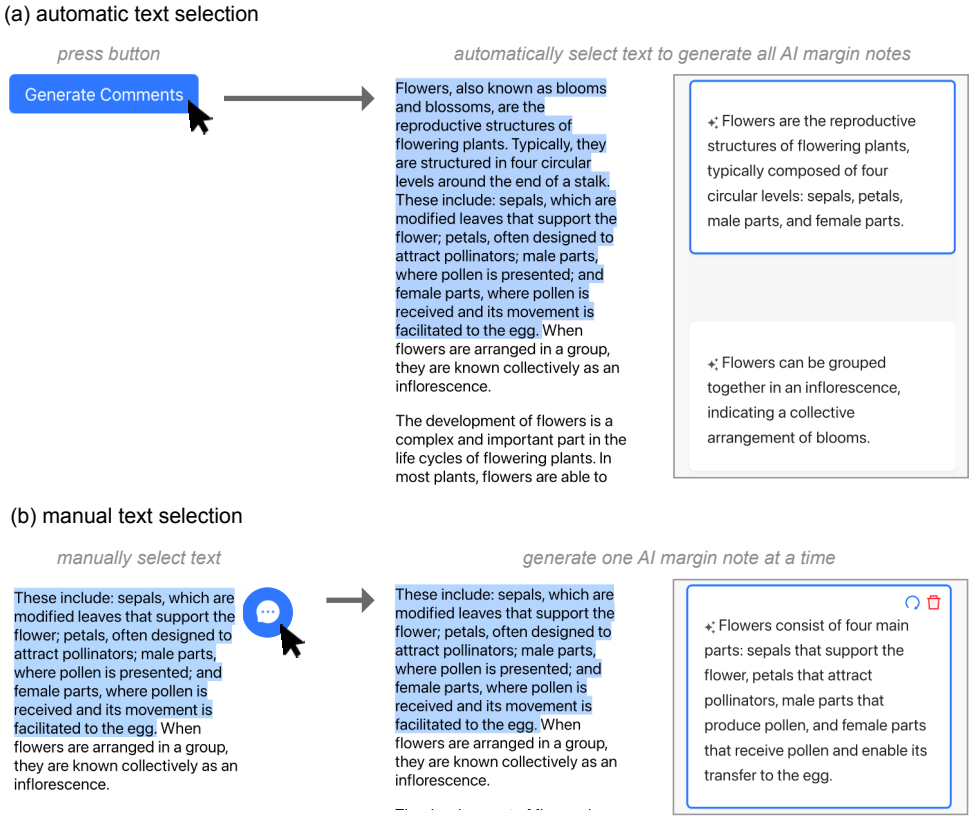

Figure 7: Techniques tested in Experiment 2: (a) automatically selecting text to generate all AI margin notes at once and (b) manually selecting text and generating AI margin notes one-by-one.

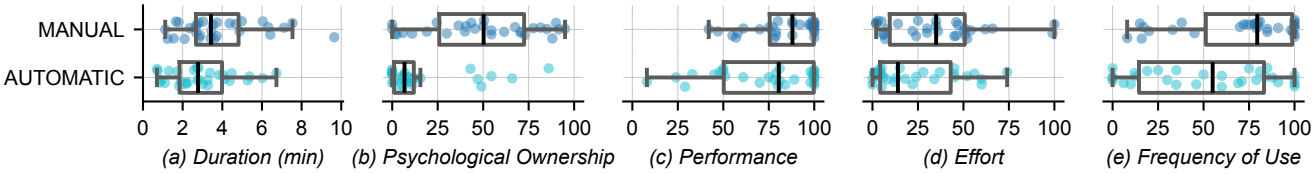

Figure 8: Experiment 2 results: (a) *Duration*, (b) *Psychological Ownership*, (c) *Performance*, (d) *Effort*, and (e) *Frequency of Use*.

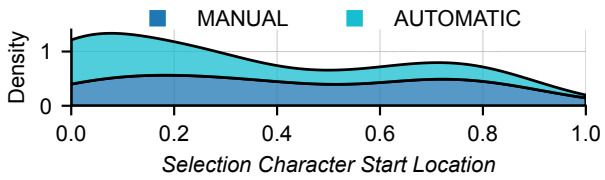

Figure 9: Distribution of selection start locations in normalized document position, shown using a kernel density estimate, where Density indicates the estimated concentration of points.

6.3.5 Selection Location and Length. The location and length of selections supports participant comments that MANUAL encouraged reading the document and enabled more control over selected text. We examined the distribution of *Selection Location* (Figure 9) and found the LLM tended to place comments at the beginning of the document with AUTOMATIC, while participants placed them more consistently throughout the document with MANUAL. We also calculated the *Selection Word Count*, finding selections were significantly shorter with MANUAL (MDN = 62, IQR = 83) compared to AUTOMATIC (MDN = 83, IQR = 90.75).

6.4 Summary

Overall, our results suggested that selecting text for an AI margin note is slower and requires more effort when done manually. However, this may be a worthwhile trade-off, as manually selecting text was associated with higher feelings of psychological ownership, better perceived performance, and was preferred by participants. This may have been due in part to the increased control over where text selections were placed, as automatic text selections tended to be placed at the beginning of the document. In this experiment, the

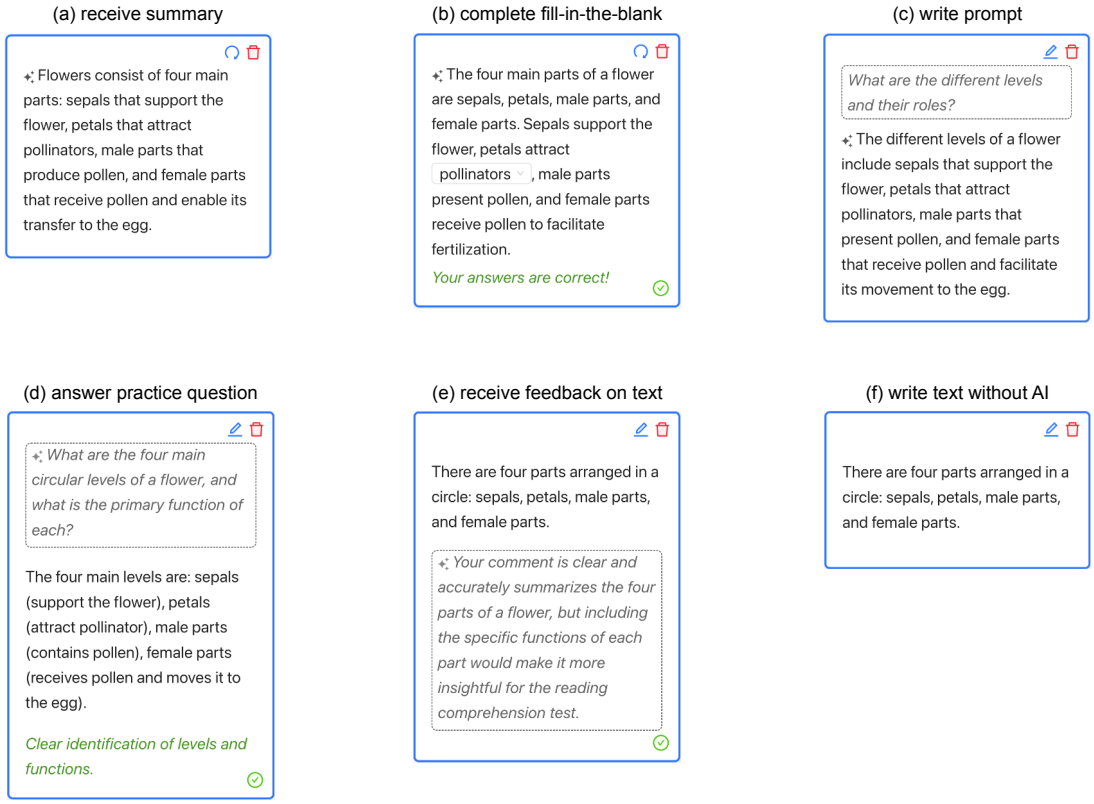

Figure 10: Techniques tested in Experiment 3: (a) receive a summary (100% AI-written text), (b) complete a fill-in-the-blank exercise, (c) write a prompt, (d) answer a practice question, (e) receive feedback on written text, and (f) write a comment without AI (100% human-written text).

resulting AI margin note was always a summary, however, there are many other types of AI margin notes that can be created that involve the human and AI in different ways.

7 AI Margin Note Techniques

To fully explore the design parameter of human and AI involvement, we implemented six techniques that fall along a continuum, ranging from AI margin notes whose text was generated entirely by the LLM, to those whose text was written entirely by the user (Figure 10). All are created by selecting text manually and pressing a blue "Comment" button, and *all are demonstrated in the supplementary video.*

7.1 Receive Summary

The technique with the least human involvement is receiving a generated summary where all of the text is written by AI (Figure 10a). This represents a common way LLMs are used in document readers and is identical to the manual text selection technique presented in Experiment 2.

7.2 Complete Fill-in-the-Blank Exercise

A technique with slightly more human involvement is receiving a generated summary, but with some keywords omitted to resemble a fill-in-the-blank question (Figure 10b). Like the previous technique, the LLM produces a summary for the text selected by the user, shown in black text and prepended by a black 'sparkle' icon. Each summary contains one or two 'blanks,' displayed using dropdowns containing three options. The user must select the correct answer from the provided options. Underneath, the comment displays feedback to notify the user whether their responses are correct (shown in green with a check mark) or incorrect (shown in red with an 'X'). The user can delete the comment or regenerate it. Fill-in-the-blank questions are a common way of assessing reading comprehension [46, 58] as they require readers to infer the contents of missing words based on their own understanding of the text [32].

7.3 Write Prompt

A technique with some human and some AI involvement is writing a custom prompt for the LLM to respond to (Figure 10c). Writing prompts is the primary way users interact with current chat-based tools and allows the user to achieve goals that go beyond summarization, such as learning additional information about the text or simplifying it. Formulating a prompt also requires users to form goals and subtasks, which may encourage them to reflect on their own understanding of the text [57]. This is identical to the AI margin note technique described in Experiment 1.

Table 3: Experiment 3 demographics.

Gender		Age		Education	
Men	13	18-24	1	High School	7
Women	19	25-34	2	Some University (no credit)	2
		35-44	7	Technical Degree	5
		45-54	13	Bachelor's Degree	14
		55-64	4	Master's Degree	4
		65-74	4		
		75+	1		

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	6	Daily	1	Daily	13	Daily	3
Weekly	14	Weekly	6	Weekly	11	Weekly	12
Monthly	8	Monthly	4	Monthly	5	Monthly	3
Less than Monthly	3	Less than Monthly	8	Less than Monthly	3	Less than Monthly	10
Never	1	Never	12			Never	4

7.4 Answer Practice Question

Another technique with some involvement from both the human and the AI is answering a practice reading comprehension question that was generated by the LLM, which has been shown to help with learning [41, 62]. The LLM produces a practice short-answer question based on the selected text (Figure 10d), prepended with a grey 'sparkle' icon and shown using grey italicized text with a dashed border. The user types in their answer in a text box underneath and presses a blue "Confirm" button to submit their response. The submitted response is then shown in black text. After a few seconds, the user receives feedback from the LLM about their answer: correct responses are shown in green and with a check mark underneath the response, and incorrect responses are shown in red with an 'X.' All responses have to be correct. The user can edit their responses to correct mistakes (and will receive new feedback accordingly), or can delete their comment.

7.5 Receive Feedback on Written Text

A technique that has a lot of human involvement is requiring users to write their own comment text, with LLM-generated feedback and suggestions on how their comment can be improved (Figure 10e). This can be beneficial, as many learners struggle to take effective notes [10]. Prior work suggests that when writing, AI-generated feedback and suggestions can help users see alternative perspectives and encourage reflection [4, 18], which could be helpful in the context of note-taking. The user can type within the provided text box and press "Confirm." The comment text is shown as black text. After a few seconds, the AI assistant provides feedback on the text, prepended with a grey 'sparkle' icon and shown as grey, italicized text with a dashed border. A green check mark is shown if their comment text is 'good,' and a red 'X' is shown if it is not. All comment text has to be classified as good. The user can edit their comment to receive updated feedback and can delete it.

7.6 Write Text without AI

The technique with the most human involvement is requiring the user to write their own comment, without any AI assistance (Figure 10f). This baseline resembles the behaviour of existing commenting systems, where all of the text is written by the user. The user can type within the provided text box and press "Confirm." The comment is shown as black text. The user can edit or delete the comment.

8 Experiment 3: Human and AI Involvement

The previous experiment found that manual text selection is preferred, but it did not explore different ways to generate AI margin notes after selection. The goal of this experiment is to investigate the effect of human and AI involvement when generating the AI margin note itself. Participants read six documents, one for each technique described above: receiving a summary (SUMMARY), completing a fill-in-the-blank exercise (BLANK), writing a prompt (PROMPT), answering a practice question (QUESTION), receiving feedback on written text (FEEDBACK), and writing without any AI assistance (NONE). The experiment was longer, approximately 90 minutes.

8.1 Participants

We recruited 36 new participants on Prolific, using the same inclusion criteria as the previous experiments. Four (11%) were excluded for experiencing technical issues, leaving 32 valid responses (Table 3). All but one were proficient at reading in English. Participants received $25 total, with an additional $5 bonus if their total reading comprehension score was within the top 25%.

8.2 Results

Where applicable, we use Friedman omnibus tests and Wilcoxon signed-rank post hoc tests, with Holm's corrections for multiple comparisons; and Spearman's correlations. *Statistical test details are shown in Table A.3.*

8.2.1 Reading Comprehension. Contrary to what prior work suggests, techniques with more human involvement were not associated with higher reading comprehension scores. Specifically, the

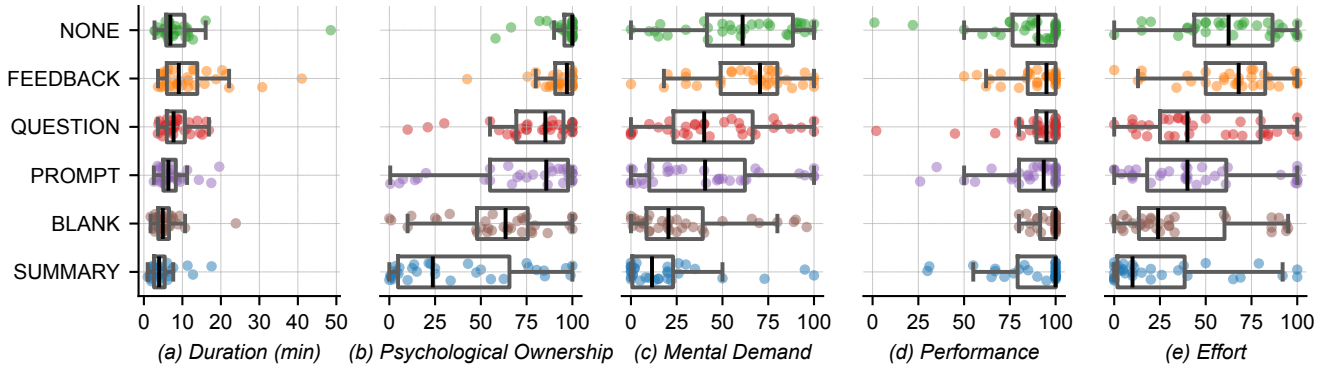

Figure 11: Experiment 3 results: (a) *Duration*, (b) *Psychological Ownership*, (c) *Mental Demand*, (d) *Performance*, and (e) *Effort*.

differences between SUMMARY (MDN = 4, IQR = 2.25) and all other conditions (all MDN = 5, IQR = 2) are marginal ($p = .054$), and not statistically significant.

8.2.2 Duration. Participants were faster when they used techniques that did not require as much human involvement (Figure 11a). A significant effect of CONDITION on *Duration* revealed that SUMMARY (MDN = 3.96, IQR = 2.78) and BLANK (MDN = 4.92, IQR = 2.80) were faster than all other techniques.

8.2.3 Psychological Ownership. Generally, participants felt more psychological ownership for techniques that required more human involvement (Figure 11b). A significant effect of CONDITION on *Psychological Ownership* and post hoc tests revealed that NONE (MDN = 100, IQR = 4.38) and FEEDBACK (MDN = 97, IQR = 9.34) were associated with the highest *Psychological Ownership*, and SUMMARY (MDN = 23.75, IQR = 60.88) with the lowest.

8.2.4 Task Workload. Generally, techniques that required more human involvement were more mentally demanding, associated with poorer perceived performance, and more effortful. For *Mental Demand*, there was a significant effect of CONDITION (Figure 11c) and post hoc tests showed that NONE (MDN = 61, IQR = 47) and FEEDBACK (MDN = 70.5, IQR = 31) were associated with higher scores than PROMPT (MDN = 40.5, IQR = 52.25), BLANK (MDN = 20.5, IQR = 30.75), and SUMMARY (MDN = 11.5, IQR = 22.25). There was also a significant effect of CONDITION on *Performance* (Figure 11d), with post hoc tests revealing that participants generally felt like they did better at the task with BLANK (MDN = 100, IQR = 8.5) than NONE (MDN = 90.5, IQR = 23.5), FEEDBACK (MDN = 95, IQR = 15.25), and SUMMARY (MDN = 100, IQR = 20.75). For *Effort*, a significant effect of CONDITION (Figure 11e) and post hoc tests showed that NONE (MDN = 62.5, IQR = 42.75) and FEEDBACK (MDN = 68, IQR = 32.5) had the highest scores, and SUMMARY (MDN = 10, IQR = 36.75) had the lowest. We did not observe significant differences between the different techniques for *Physical Demand*, *Temporal Demand*, and *Frustration*.

8.2.5 Preferences. There were no differences in *Frequency of Use*, with all medians being 50 or greater, suggesting that participants would like to use all techniques. For *Overall Ranking*, the Condorcet voting method revealed that overall, BLANK was ranked first, followed by QUESTION, SUMMARY, PROMPT, FEEDBACK, and NONE. This *Overall Ranking* suggests that participants generally preferred techniques with more AI involvement (BLANK, QUESTION, SUMMARY, and PROMPT) over those with more human involvement (FEEDBACK and NONE), for example: *"the more involved AI was, the more I appreciated the help"* (P15).

Considering the *Ranking* scores, 21 (66%) ranked BLANK within the top 3, and 20 (62%) placed QUESTION within the top 3. These two techniques were generally valued as they *"had the greatest pedagogical heft"* (P7) and *"actively engage memory and reinforce understanding"* (P4). Nineteen (59%) placed SUMMARY within the top 3, however, eight (25%) gave it a rank of 6, suggesting that opinions were more divided for SUMMARY. Those that ranked SUMMARY highly generally valued it for *"[making] work so much easier"* (P9), however, some noted how they *"want to do things [and] feel accomplished"* (P30), goals that were not as well-supported with SUMMARY. Fifteen (47%) placed PROMPT within the top 3 as it was *"simple and easy"* (P23) and *"encouraged curiosity"* (P4). In contrast, seventeen (53%) ranked FEEDBACK within the bottom 3, and 19 (59%) ranked NONE within the bottom 3. Both techniques required *"too much effort"* (P20), which could be *"draining and mind-absorbing"* (P25).

8.2.6 Text Similarity. We hypothesized that techniques with more human involvement could discourage readers from taking near-verbatim notes [10]. To determine how similar the text typed by the participant was to the selected text, we used Google's Universal Sentence Encoder [14] to calculate semantic *Text Similarity* (0-1 range where 1 means identical).[6] As SUMMARY and BLANK do not require the reader to write any text, we compare the selected text to the text that was generated by the LLM. Overall, our results suggested that techniques that required writing from the participants were associated with lower *Text Similarity* than those that did not. A significant effect of CONDITION on *Text Similarity* and post hoc tests showed that SUMMARY (MDN = .78, IQR = .14) and BLANK (MDN = .8, IQR = .15) had higher *Text Similarity* scores than FEEDBACK (MDN = .62, IQR = .28), QUESTION (MDN = .49, IQR = .28), PROMPT (MDN = .49, IQR = .3), and NONE (MDN = .58, IQR = .23).

[6]Note that we also calculated *Text Similarity* using typed text concatenated with LLM-generated questions and responses for QUESTION and PROMPT and found similar results.

8.3 Summary

Our results only suggested marginal differences in reading comprehension, even though many techniques were more mentally demanding and effortful. Participants generally preferred techniques with more AI involvement, even though these techniques were associated with less psychological ownership.

9 Discussion

We summarize principle findings and introduce design implications for each design parameter, then consider how aspects of AI margin notes could transfer to other contexts and acknowledge limitations in our approach.

9.1 Integration

Participants had a strong preference for AI margin notes citing several perceived benefits associated with increased integration: (1) convenient reference to specific concepts in the text when formulating prompts; (2) no need to shift attention to a separate chat interface, and (3) quick access to prior LLM responses.

It is well-known that natural language prompts must be explicit and specific [19, 44], but people find this challenging to do [19, 57]. Reading and note-taking are already cognitively-demanding activities [52], so the additional task of formulating explicit and specific prompts may be too much [25]. However, using deictic language is often easier than fully articulated descriptions [9, 22], which likely applies to prompts as well. For example, DirectGPT [44] encouraged prompts with deictic language by using direct manipulation to point at deictic references. AI margin notes also encourage deictic language by associating each prompt with specific text selected using a standard direct manipulation interaction. We believe this frees users to focus more on metacognitive tasks like identifying parts of the document where they require help [57], rather than the nuances of language [67].

Design Implications. The primary implication is to **adopt an integrated AI margin note approach within document reader software**. A secondary design implication is to **devise ways to prompt with more deictic language, less interface switching, and more ways of retrieving previous responses**. For example, allowing readers to place prompts and responses at arbitrary locations independent of text selection, like the 'sticky note' feature in most document reader software.

9.2 Selection Automation

Selecting text manually was slower and required more effort, but some participants noted this encouraged them to read the text more. Furthermore, *manual selection increased psychological ownership with participants feeling like they were more effective.* This was supported by shorter manual text selections that were placed more uniformly throughout the document. Together, these effects likely contributed to the *strong preferences participants had for manually-created AI margin notes*, with most describing how they *valued the increased control enabled with manual text selections.*

When considering selection automation and human and AI involvement, we observe similar results regarding duration, effort, and psychological ownership. Yet user preferences diverged, with participants preferring more manual control over text selections but more AI involvement within the actual comment. Together, these findings provide additional insight into factors that are more important to users when reading and taking notes. Based on prior work on text highlighting [33, 66], we hypothesize that manual selection forces the reader to identify text they wish to learn more about, which requires them to recognize gaps in their own knowledge or understanding of the text [57]. An LLM does not know what these gaps are, making it less capable of automatically positioning AI margin notes in ways that align with the reader's needs. Identifying the right text to select is especially important, as the selected text also acts as contextual information for content that is produced and displayed within the comment. As such, *selection automation may be a task that is seen to be less compatible with more automation and AI involvement.* However, once text is manually selected, there is more potential for AI involvement in the creation of the margin note itself.

Design Implications. The primary implication is to **prioritize manual text selection when creating AI margin notes**. A secondary implication is to **devise ways of encouraging more control over automatically-placed AI margin notes** when such capabilities are necessary. For example, readers often struggle to identify the most important information [21], and systems like Paper Plain [3] suggest that automatically-generated summaries can help readers understand complex documents. Automatically-placed AI margin notes may have a similar effect, especially when designed in ways that give users more control. One idea is presenting automatically-placed AI margin notes as suggested comments that the reader must manually "confirm" to save. This may improve feelings of psychological ownership [40], ensure that the AI margin notes are placed more consistently throughout the document, and encourage users to read the text more closely.

9.3 Human and AI Involvement

Our results suggest that *AI margin note techniques have their own strengths and weaknesses.* For example, techniques with more human involvement, like receiving feedback about written text, required more mental demand and effort, but were associated with higher feelings of psychological ownership. Yet, participants preferred techniques with more AI involvement, even fully automated summary notes. This contrasts with the desire to manually select text, but it aligns with Kreijkes et al.'s [39] findings for LLM-assisted note-taking while reading. In a different note-taking context involving a live video lecture, Chen et al. [16] compared using an LLM to automatically organize generated summaries and transcripts to manual organization by participants, and also found participants preferred more AI involvement.

However, both Kreijkes et al. [39] and Chen et al. [16] also found that note-taking techniques with more human involvement increased comprehension. This aligns with work suggesting that increased cognitive engagement can increase learning [5–7], so we are surprised that more cognitively demanding AI margin note techniques were not associated with higher reading comprehension scores. To confirm this was not due to our experimental protocol, we conducted other exploratory experiments with different participants and key variations. We ran experiments with 24 hour gaps

between the reading and testing stages to ensure that the documents were not simply remembered in short-term memory, but we did not observe any significant differences. Participants self-declared their knowledge of the topics discussed in each document, so we tried omitting participants with higher background knowledge, and found similar results. Another possibility is a ceiling effect due to a smaller range of possible comprehension scores (0-6). To rule this out, we conducted another experiment using documents and questions from Guidroz et al. [27],[7] with more granular (0-12) scores. These documents were also at a more difficult university graduate level. Yet again, we found similar results, with no significant differences in comprehension between techniques.

Design Implications. We believe that less cognitively engaging techniques with more AI involvement, like receiving generated summaries, may not be as detrimental to reading comprehension as one may assume. However, our results did suggest that summaries, which featured no human involvement, were not as preferred as other techniques with a little more human involvement, suggesting that readers do not want to offload all responsibilities to AI. Therefore, our primary implication for design is to **prioritize AI margin note techniques that balance human and AI involvement**. Given the wide range of perceived benefits and trade-offs of each technique, a secondary design implication is to **provide multiple options of AI margin note techniques that vary in human and AI involvement** to better suit reader preferences and goals. This idea of human and AI involvement could also extend to chat-based interfaces, for example, through techniques that scaffold prompts to elicit more input from the user [57].

9.4 Other AI Margin Note Designs

Receiving generated summaries with fill-in-the-blank exercises was the most preferred technique. It was perceived as not too mentally demanding or effortful, still associated with moderate feelings of psychological ownership, and participants felt like they performed well when creating them. An exciting avenue for future work is to explore other AI margin note designs that focus on these aspects. For example, using digital ink and sketches [54] to prompt LLMs [64]. Interactive visual elements, like simulations [28] and charts [43] could be generated by an LLM and integrated as AI margin notes. For example, if the reader selects text that describes different parts of flowering plants, a diagram could be generated with fill-in-the-blank labels for different parts of a flower.

9.5 Adapting AI Margin Notes to Other Contexts

Beyond document reader software, integrating prompting with text selections could prove to be useful in other textual domains. Consider a code editor like VS Code, where code can be selected and an in-line prompt triggered. The explicit text context and ability to use deictic language is similar to AI margin notes, but the prompts and responses are moved to the side chat panel, or not saved at all. Adopting the AI margin note convention of persisting the note in a margin near the associated text (e.g., in the gutter with line numbers) could make prompts easier to reuse and make AI use more transparent for collaborators.

[7]We emailed the authors and received permission to reuse their materials.

The AI margin note approach could be applied to domains other than text. Clicking on a UI element and prompting using deictic language could be the foundation for general software help-seeking [22]. This could be automated based on past user behaviour, where multiple AI notes are overlaid on an application interface, each pointing to a part of the UI that typically requires explanation.

9.6 Limitations

Our results, especially results related to reading comprehension, may not hold after extended long-term use. For example, LLMs may hallucinate and produce incorrect AI margin notes, and people may come to over-rely by studying incorrect notes [61]. Techniques with low human involvement may make students 'lazy' over time and impact their ability to take notes in situations where LLMs are unavailable. Assessing the impact of AI margin note techniques in a wider range of educational settings is an important direction for future work.

Some aspects of our experimental design may be lacking in ecological validity. Notably, we chose to keep the number of AI margin notes and chat responses constant across techniques, and participants could only interact with one technique at a time. We made these decisions for increased control. In a real system, these restrictions would not exist: users could create as many AI margin notes and in whatever way they wish.

Our idea of integrating prompting into comments, which are inherently smaller in size, means that there are limits of how much content can be displayed and how many rounds of interactions are possible: with too much content, each AI margin note could become too 'chat-like,' nullifying any perceived benefits over chat-based interfaces. Future work could explore ways to adapt or extend AI margin notes when considering this space constraint.

10 Conclusion

We propose and explore the design of "AI margin notes" that leverage the commenting feature of document reader software to provide LLM capabilities in a way that is more integrated into document text. Three experiments evaluated variations from different design parameters: integration, selection automation, and human and AI involvement, and overall, participants valued having integrated AI margin notes and creating them manually. AI margin note techniques that involved the human and AI to different degrees were valued for different reasons, suggesting that document reader software should provide multiple variations to support different user goals and preferences. Our work adds more evidence that chat-based interfaces are not the only way of interacting with LLMs, and that increased integration with document text is beneficial, especially when they are created manually and the trade-offs of human and AI involvement are considered.

Acknowledgments

This work was made possible by NSERC Discovery Grant 2024-03827.

References

[1] Annette Adler, Anuj Gujar, Beverly L. Harrison, Kenton O'Hara, and Abigail Sellen. 1998. A diary study of work-related reading: design implications for digital reading devices. In *Proceedings of the SIGCHI Conference on Human Factors in*

Computing Systems (Los Angeles, California, USA) *(CHI '98)*. ACM Press/Addison-Wesley Publishing Co., USA, 241–248. doi:10.1145/274644.274679

[2] Adobe. 2025. Generative AI in Adobe Document Cloud Applications. https://helpx.adobe.com/ca/acrobat/using/generative-ai.html. [Accessed 20-08-2025].

[3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 74 (Sept. 2023), 38 pages. doi:10.1145/3589955

[4] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1049, 18 pages. doi:10.1145/3613904.3642406

[5] Elizabeth L Bjork and Robert A Bjork. 2011. Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* 2, 59-68 (2011), 56–64.

[6] Robert A Bjork. 1994. Institutional Impediments to Effective Training. *Learning, Remembering, Believing: Enhancing Human Performance* (1994), 295–306.

[7] Robert A Bjork. 1994. Memory and Metamemory Considerations in the Training of Human Beings. *Metacognition: Knowing about Knowing* 185, 7.2 (1994), 185–205.

[8] Melanie Ramdarshan Bold and Kiri L Wagstaff. 2017. Marginalia in the Digital Age: Are Digital Reading Devices Meeting the Needs of Today's Readers? *Library & Information Science Research* 39, 1 (2017), 16–22.

[9] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[10] Burke H. Bretzing and Raymond W. Kulhavy. 1979. Notetaking and Depth of Processing. *Contemporary Educational Psychology* 4, 2 (1979), 145–153. doi:10.1016/0361-476X(79)90069-9

[11] Burke H Bretzing and Raymond W Kulhavy. 1981. Note-Taking and Passage Style. *Journal of Educational Psychology* 73, 2 (1981), 242.

[12] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[13] Avner Caspi and Ina Blau. 2011. Collaboration and Psychological Ownership: How does the Tension between the Two Influence Perceived Learning? *Social Psychology of Education* 14 (2011), 283–298.

[14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 169–174.

[15] ChatPDF. 2023. ChatPDF AI | Chat with any PDF | Free — chatpdf.com. https://www.chatpdf.com/. [Accessed 29-08-2025].

[16] Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2025. More AI Assistance Reduces Cognitive Engagement: Examining the AI Assistance Dilemma in AI-Supported Note-Taking. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW451 (Oct. 2025), 29 pages. doi:10.1145/3757632

[17] Fergus I.M. Craik and Robert S. Lockhart. 1972. Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior* 11, 6 (1972), 671–684. doi:10.1016/S0022-5371(72)80001-X

[18] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. doi:10.1145/3526113.3545672

[19] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. doi:10.1145/3544548.3580969

[20] Ruiqi Deng, Maoli Jiang, Xinlu Yu, Yuyan Lu, and Shasha Liu. 2025. Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-Analysis of Experimental Studies. *Computers & Education* 227 (2025), 105224.

[21] Robert L Fowler and Anne S Barker. 1974. Effectiveness of Highlighting for Retention of Text Material. *Journal of Applied Psychology* 59, 3 (1974), 358–364. doi:10.1037/h0036750

[22] C. Ailie Fraser, Julia M. Markel, N. James Basa, Mira Dontcheva, and Scott Klemmer. 2020. ReMap: Lowering the Barrier to Help-Seeking with Multimodal Search. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 979–986. doi:10.1145/3379337.3415592

[23] Google. 2025. Google NotebookLM | AI Research Tool & Thinking Partner — notebooklm.google. https://notebooklm.google. [Accessed 20-08-2025].

[24] Scott C Griffith. 1990. Cooperative Learning Techniques in the Classroom. *Journal of Experiential Education* 13, 2 (1990), 41–44.

[25] Axel Grund, Stefan Fries, Matthias Nückles, Alexander Renkl, and Julian Roelle. 2024. When is Learning "Effortful"? Scrutinizing the Concept of Mental Effort in Cognitively Oriented Research from a Motivational Perspective. *Educational Psychology Review* 36, 1 (2024), 11.

[26] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 898, 22 pages. doi:10.1145/3613904.3642699

[27] Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, et al. 2025. LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load. *arXiv preprint arXiv:2505.01980* (2025).

[28] Aditya Gunturu, Yi Wen, Nandi Zhang, Jarin Thundathil, Rubaiat Habib Kazi, and Ryo Suzuki. 2024. Augmented Physics: Creating Interactive and Embedded Physics Simulations from Static Textbook Diagrams. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 144, 12 pages. doi:10.1145/3654777.3676392

[29] Beverly L Harrison, Hiroshi Ishii, Kim J Vicente, and William AS Buxton. 1995. Transparent Layered User Interfaces: An Evaluation of a Display Design to Enhance Focused and Divided Attention. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 317–324.

[30] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[31] Heather Joanna Jackson. 2001. *Marginalia: Readers Writing in Books*. Yale University Press.

[32] J Jonz. 1994. The Effects of Textual Cohesion and Prior Knowledge on Native and Nonnative Cloze Test Scores. *Cloze and Coherence* (1994), 269–285.

[33] Nikhita Joshi and Daniel Vogel. 2024. Constrained Highlighting in a Document Reader can Improve Reading Comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–10.

[34] Nikhita Joshi and Daniel Vogel. 2025. Interaction Techniques that Encourage Longer Prompts Can Improve Psychological Ownership when Writing with AI. arXiv:2507.03670 [cs.HC] https://arxiv.org/abs/2507.03670

[35] Nikhita Joshi and Daniel Vogel. 2025. Writing with AI Lowers Psychological Ownership, but Longer Prompts Can Help. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 72, 17 pages. doi:10.1145/3719160.3736608

[36] Kenneth A Kiewra. 1985. Investigating Notetaking and Review: A Depth of Processing Alternative. *Educational psychologist* 20, 1 (1985), 23–32.

[37] Kenneth A Kiewra. 1989. A Review of Note-Taking: The Encoding-Storage Paradigm and Beyond. *Educational Psychology Review* 1, 2 (1989), 147–172.

[38] Keiichi Kobayashi. 2005. What Limits the Encoding Effect of Note-Taking? A Meta-Analytic Examination. *Contemporary Educational Psychology* 30, 2 (2005), 242–262.

[39] Pia Kreijkes, Viktor Kewenig, Martina Kuvalja, Mina Lee, Sylvia Vitello, Jake M Hofman, Abigail Sellen, Sean Rintel, Daniel G Goldstein, David M Rothschild, Lev Tankelevitch, and Tim Oates. 2025. Effects of LLM Use and Note-Taking on Reading Comprehension and Memory: A Randomised Experiment in Secondary Schools. *Available at SSRN* (2025).

[40] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship. In *Proceedings of Mensch Und Computer 2022* (Darmstadt, Germany) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA, 192–208. doi:10.1145/3543758.3543947

[41] Ming Liu, Jingxu Zhang, Lucy Michael Nyagoga, and Li Liu. 2023. Student-AI Question Cocreation for Enhancing Reading Comprehension. *IEEE Transactions on Learning Technologies* 17 (2023), 815–826.

[42] Catherine C Marshall. 1997. Annotation: From Paper Books to the Digital Library. In *Proceedings of the Second ACM International Conference on Digital Libraries*. 131–140.

[43] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2023. Charagraph: Interactive Generation of Charts for Realtime Annotation of Data-Rich Paragraphs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 146, 18 pages. doi:10.1145/3544548.3581091

[44] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 975, 16 pages. doi:10.1145/3613904.3642462

[45] Liam Melin-Higgins. 2024. AI in Document Interaction: An Interaction Design Approach to Enhancing Reading Practices.

[46] Jakob Nielsen. 2011. Cloze Test for Reading Comprehension. https://www.nngroup.com/articles/cloze-test-reading-comprehension/. [Accessed 02-12-2025].

[47] Jakob Nielsen. 2015. Legibility, Readability, and Comprehension: Making Users Read Your Words. https://www.nngroup.com/articles/legibility-readability-comprehension/. [Accessed 08-12-2024].

[48] NoteGPT. 2025. NoteGPT – Your All-in-One AI Learning Assistant. Summarize, Chat & Write – Fast & Free. — notegpt.io. https://notegpt.io. [Accessed 20-08-2025].

[49] Thierry Olive and Marie-Laure Barbier. 2017. Processing Time and Cognitive Effort of Longhand Note Taking when Reading and Summarizing a Structured or Linear Text. *Written Communication* 34, 2 (2017), 224–246.

[50] Jon L. Pierce, Tatiana Kostova, and Kurt T. Dirks. 2001. Toward a Theory of Psychological Ownership in Organizations. *The Academy of Management Review* 26, 2 (2001), 298–310. doi:10.2307/259124

[51] Jon L. Pierce, Tatiana Kostova, and Kurt T. Dirks. 2003. The State of Psychological Ownership: Integrating and Extending a Century of Research. *Review of General Psychology* 7, 1 (2003), 84–107. doi:10.1037/1089-2680.7.1.84

[52] Annie Piolat, Thierry Olive, and Ronald T Kellogg. 2005. Cognitive Effort during Note Taking. *Applied Cognitive Psychology* 19, 3 (2005), 291–312.

[53] Carol Porter-O'Donnell. 2004. Beyond the Yellow Highlighter: Teaching Annotation Skills to Improve Reading Comprehension. *English Journal* 93, Secondary Readers Reading Successfully (2004), 82–89.

[54] Hugo Romat, Emmanuel Pietriga, Nathalie Henry-Riche, Ken Hinckley, and Caroline Appert. 2019. Spaceink: Making Space for In-Context Annotations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 871–882.

[55] Christopher A. Sanchez and Jennifer Wiley. 2009. To Scroll or Not to Scroll: Scrolling, Working Memory Capacity, and Comprehending Complex Texts. *Human Factors* 51, 5 (2009), 730–738. doi:10.1177/0018720809352788 PMID: 20196297.

[56] Elizabeth G Soslau and Deborah S Yost. 2007. Urban Service-Learning: An Authentic Teaching Strategy to Deliver a Standards-Driven Curriculum. *Journal of Experiential Education* 30, 1 (2007), 36–53.

[57] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 680, 24 pages. doi:10.1145/3613904.3642902

[58] Wilson L. Taylor. 1953. "Cloze Procedure"': A New Tool for Measuring Readability. *Journalism Quarterly* 30, 4 (1953), 415–433. doi:10.1177/107769905303000401

[59] Sherman W Tyler, Paula T Hertel, Marvin C McCallum, and Henry C Ellis. 1979. Cognitive Effort and Memory. *Journal of Experimental Psychology: Human Learning and Memory* 5, 6 (1979), 607.

[60] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 38 (March 2022), 56 pages. doi:10.1145/3502222

[61] Jin Wang and Wenxiang Fan. 2025. The Effect of ChatGPT on Students' Learning Performance, Learning Perception, and Higher-Order Thinking: Insights from a Meta-Analysis. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–21.

[62] Xizhe Wang, Yihua Zhong, Changqin Huang, and Xiaodi Huang. 2024. ChatPRCS: A Personalized Support System for English Reading Comprehension Based on ChatGPT. *IEEE Transactions on Learning Technologies* 17 (2024), 1722–1736.

[63] Joanna L Wolfe and Christine M Neuwirth. 2001. From the Margins to the Center: The Future of Annotation. *Journal of Business and Technical Communication* 15, 3 (2001), 333–371.

[64] Ryan Yen, Jian Zhao, and Daniel Vogel. 2025. Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 872, 17 pages. doi:10.1145/3706598.3713822

[65] H Peyton Young. 1988. Condorcet's Theory of Voting. *American Political Science Review* 82, 4 (1988), 1231–1244.

[66] Carole L Yue, Benjamin C Storm, Nate Kornell, and Elizabeth Ligon Bjork. 2015. Highlighting and its Relation to Distributed Study and Students' Metacognitive Beliefs. *Educational Psychology Review* 27, 1 (2015), 69–78. doi:10.1007/s10648-014-9277-z

[67] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

A Appendix

Table A.1: Experiment 1 statistical test results.

Measure	W	p	RBC
Reading Comprehension	80	.81	.06
Duration	148	.50	.16
Psychological Ownership	82.5	.40	.21
Mental Demand	149	.72	.08
Physical Demand	76	.68	.11
Temporal Demand	74.5	.63	.13
Performance	70.5	.78	.08
Effort	105.5	.32	.24
Frustration	57	.12	.40
Frequency of Use	77.5	.07	.44

Table A.2: Experiment 2 statistical test results.

Measure	W	p	RBC	
Reading Comprehension	88.5	.20	.30	
Duration	102	.006	.56	**
Psychological Ownership	9	< .001	.95	***
Mental Demand	158.5	.46	.16	
Physical Demand	31.5	.10	.48	
Temporal Demand	100	.25	.28	
Performance	52.5	.02	.58	*
Effort	84.5	.004	.61	**
Frustration	77.5	.19	.33	
Frequency of Use	63.5	.004	.64	**
Selection Word Count	1320	.003	.36	**

Table A.3: Experiment 3 statistical test results.

	Measure	Q	p	W		
	Reading Comprehension	10.87	.05	.07		
(a)	*Duration*	64.64	< .001	.40	***	
(b)	*Psychological Ownership*	74.76	< .001	.47	***	
(c)	*Mental Demand*	57.83	< .001	.36	***	
	Physical Demand	20.92	< .001	.13	***	*post hocs were n.s.*
	Temporal Demand	4.45	.49	.03		
(d)	*Performance*	16.82	.005	.10	**	
(e)	*Effort*	55.98	< .001	.35	***	
	Frustration	5.39	.37	.03		
	Frequency of Use	9.19	.10	.06		
(f)	*Text Similarity*	92.96	< .001	.58	***	

		(a) *Duration*		(b) *P. Ownership*		(c) *Mental Demand*		(d) *Performance*		(e) *Effort*		(f) *Text Similarity*	
comparisons		*p-value*		*p-value*		*p-value*		*p-value*		*p-value*		*p-value*	
NONE	FEEDBACK	.02	*	.24		.52		1		.91		.07	
NONE	QUESTION	.56		< .001	***	.06		1		.004	**	.09	
NONE	PROMPT	.45		.003	**	.02	*	1		.01	*	< .001	***
NONE	BLANK	.003	**	< .001	***	< .001	***	.005	**	< .001	***	< .001	***
NONE	SUMMARY	< .001	***	< .001	***	< .001	***	1		< .001	***	< .001	***
FEEDBACK	QUESTION	.18		.002	**	.007	**	1		.01	*	.002	**
FEEDBACK	PROMPT	.04	*	.003	**	.008	**	1		.01	*	< .001	***
FEEDBACK	BLANK	< .001	***	< .001	***	< .001	***	.03	*	.001	**	< .001	***
FEEDBACK	SUMMARY	< .001	***	< .001	***	< .001	***	1		< .001	***	< .001	***
QUESTION	PROMPT	.19		.44		.52		1		.91		.007	**
QUESTION	BLANK	.001	**	.004	**	.05		.08		.03	*	< .001	***
QUESTION	SUMMARY	< .001	***	< .001	***	< .001	***	1		.004	**	< .001	***
PROMPT	BLANK	.03	*	.24		.15		.07		.11		< .001	***
PROMPT	SUMMARY	.002	**	< .001	***	.02	*	1		.02	*	< .001	***
BLANK	SUMMARY	.14		.003	**	.15		.04	*	.04	*	.05	