Accelerating Review Efficiency: Topic-based Auto-highlighting and Grouping with LLMs

Allen Jue
Department of Computer Science
Austin, Texas, USA
mrallenjue@utexas.edu

Temitayo Awosemo
Department of Computer Science
Austin, Texas, USA
temia@utexas.edu

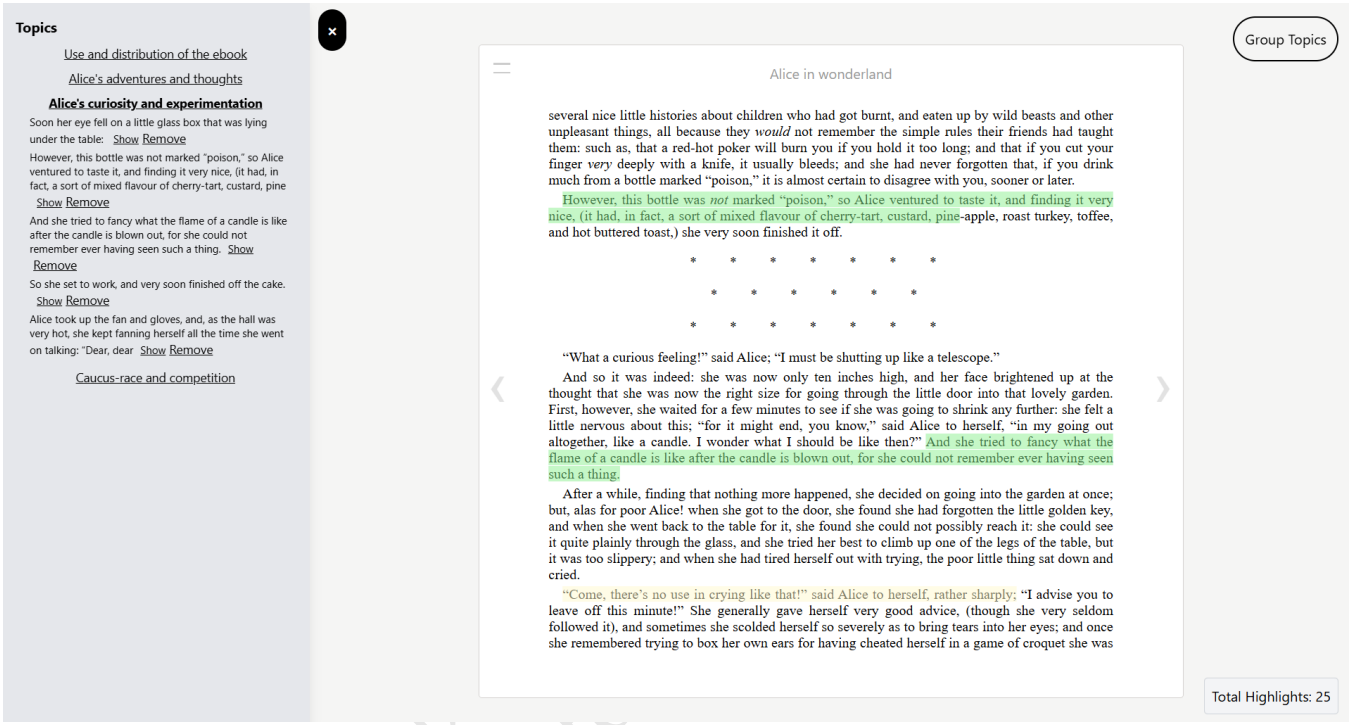

Figure 1: Low fidelity pen-and-paper mock up of topic-based auto-highlighting tool interface.

Abstract

Highlighting is a widely used note-taking and comprehension strategy, but revisiting and organizing highlights can be cumbersome. Previous research has explored methods like constrained highlighting, which improves reading comprehension by encouraging selective highlighting and learning efficient highlighting techniques [5, 8]. Other works, such as automated highlighting systems like the Semantic Reader Project, organize research papers by sections like abstract, contributions, and results [2, 10]. This paper introduces a novel thematic grouping tool that leverages large language models (LLMs) to cluster highlights into meaningful groups, offering an organized structure for text review. Prior approaches that focus on improving initial reading comprehension or information retrieval during the first read may bypass the initial cognitive engagement to properly understand the text. Our approach aims to maintain cognitive load during the initial reading while reducing future review times. We conducted a two-phase crossover study to evaluate the tool's impact on review efficiency, comprehension, and user experience using both a short ACT passage and a research paper. While our results revealed no significant differences in review speed, comprehension, or performance, qualitative findings indicate potential benefits in enhancing user engagement and organizational support. Additionally, we qualitatively observed an increased sense of confidence in understanding the text and likelihood of reviewing the text actively when using the tool.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Auto-Highlighting, Highlighting, LLMs

ACM Reference Format:

Allen Jue and Temitayo Awosemo. 2018. Accelerating Review Efficiency: Topic-based Auto-highlighting and Grouping with LLMs. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Reading is an integral part of our daily lives, and its impacts extend far beyond basic literacy. The average reading rate of an average American is 238 words per minute[1]. With just 30 minutes of reading per day, the average American can expect to read over 2.6 million words anually. Despite its prevalence, making sense of text can be difficult, and oftentimes, it is necessary to revisit a piece of text to fully comprehend its meaning—a process that can be both cognitively demanding and time-intensive.

To address this, researchers have found effective studying techniques for reading, such as non-linear note taking, active highlighting, and constrained highlighting[5, 6, 8, 14]. Highlighting is of particular interest, as it is a ubiquitous tool that students use to enhance their reading comprehension. It has been demonstrated that highlighting can benefit learning, as simply deciding what to highlight compels students to process the text at a higher conceptual level. Other benefits include the von Rerstoff effect, where pieces of text that are highlighted or visually different may be more memorable[16]. Solutions like constrained highlighting emphasize *how* to highlight and prioritize the end result of recall and comprehension. Other works, such as the Semantic Reader Project and ScentHighlights try to lower the cognitive load of reading by improving the information retrieval process [2, 10, 15].

While these approaches have many applications, these works are focused on either proper note taking strategies and efficient information retrieval during the first reading. Instead, our work is focused on maintaining the natural cognitive effort required for initial readings while accelerating subsequent review tasks. In this paper, we explore an alternative approach of reducing reading review time by leveraging large language models (LLMs) to generate automated highlights and topic groupings. We hypothesize that such a tool can decrease review time while preserving recall and comprehension.

Contributions:

- We create a text-highlighting prototype that integrates GPT-4o mini to generate automatic topic groups and highlights.
- We evaluate the effectiveness of this system and find no significant evidence that LLM-based topic grouping and highlighting reduce review time, though they maintain comparable comprehension and recall.

2 Method

2.1 Tool Development

Our tool leverages the open-source repository, Epub.js, which is integrated with a React front end to efficiently render electronic books in ePub files and generate an intuitive and interactive reading interface. We focused on maintaining a simple highlighting mechanism, where users initiate text selection by pressing and dragging the cursor, and highlighting ends upon releasing the cursor. We added a sidebar that could be expanded, which contains the existing highlights, and users can click `show` to maneuver to the highlight or `remove` to remove the highlight. When users group topics, the grouped topic names appear in the sidebar and can be expanded to show which highlights are corresponding to which LLM-generated topic.

To interface with GPT-4o mini, we utilized a client-server architecture that facilitates real-time topic generation and annotation management. The server leverages the OpenAI API to communicate with the GPT model, and it takes care of request handling, model invocation, and response processing. The client-side implementation, built with React, handles user interaction and communication with the server. The system begins with a structured prompt that explicitly instructs the GPT model to organize the sentences into descriptive and distinct topics. This step required careful prompt engineering to ensure the responses were both accurate and easy to parse. Choosing the appropriate model was a critical part of the integration. After testing a few models, including GPT-3.5-turbo, we selected GPT-4o-mini for its combination of speed and its capability to handle clustering tasks with high accuracy. The output from GPT is returned as a JSON object, which the client processes to display the topics in the expandable sidebar to the user.

2.2 Study Design

The evaluation of the tool's effectiveness was conducted through a two-phase crossover study design. This design aimed to assess review efficiency, low-level understanding, and recall, while reducing carryover effect from participant tool familiarity.

2.2.1 Phase One: Highlighting. Participants were randomly assigned to two groups, and they were provided with two reading passages: a short narrative excerpt and a longer technical article. The inclusion of both short and long passages aimed to capture differences in reading behavior and comprehension across different text types. The short passage that participants read is an excerpt from *The Men of Brewster Place*, which is an standardized ACT reading test passage [12]. The long passage is an excerpt from *Beyond Being There*[7]. It was chosen because although it is research paper, it is generally accessible to read without prior technical experience. The short passage is always read before the long passage to control for reader fatigue and order effects. By standardizing the sequence, the experiment minimizes the potential for participants to be disproportionately influenced by the length of the passage or the cumulative effort of reading multiple texts.

In Group 1, participants read the short passage using the thematic grouping tool and the long without it. Conversely, in Group 2, participants read the short passage without the tool and the long passage with the thematic grouping tool. Participants were instructed to read and highlight key points based on their judgment, with no time constraints.

2.2.2 Phase Two: Review and Assessment. After a 10-day interval, participants were tasked with reviewing their highlights. The delay was introduced to mimic real-world scenarios where time elapses

between initial reading and review, allowing for a natural forgetting curve and creating a genuine need to revisit the material. Both groups reviewed the two passages under the conditions set in Phase One. Comprehension and recall were measured using a structured questionnaire comprising of multiple-choice questions designed to probe for the understanding of central themes in both passage. For the short passage, we utilize the existing ACT questions, which are recall and comprehension based, ranging from definitions to inferring themes. For the long passage, questions were written to emulate the difficulty of the ACT questions.

2.3 Data Collection

The study employed a mixed-methods approach to data collection, using both quantitative metrics and qualitative feedback to evaluate the tool's impact comprehensively.

2.3.1 Quantitative Metrics. In phase 1, participants were timed to measure the duration of their reading sessions. Additional quantitative data includes the number of highlights and the frequency of utilizing the tool for grouping actions. In phase 2, participants were timed to measure the duration of their review sessions. Participants were also timed for how long they took the for the assigned quizzes. Their quiz scores were recorded to test for recall, comprehension, and task efficiency.

2.3.2 Qualitative Feedback. To complement the quantitative metrics, qualitative insights were captured through behavioral observations and post-experiment surveys. Participants were asked to share their impressions of the prototype, focusing on its usability, effectiveness, and ease of use. Observations of user behavior during the task, such as encountering bugs, confusion, and usage patterns were also recorded to identify potential areas for improvement.

3 Results

At the time of writing, we have tested the model on the researchers and undergraduate students (N=4).

3.1 User Performance

3.1.1 Phase One. User performance for the most part was intuitive for phase one. For the longer text, users took a longer time to read and usually utilized more highlights ($\mu = 30.67, \sigma = 38.59$), compared to the shorter text ($\mu = 8.67, \sigma = 10.78579$). Of interest is the variability in the amount of highlights. This high degree of variability possibly stems from a lack of proper training with highlighting. A small number of additional highlights were included due to learning how to utilize the tool's highlighting feature.

3.1.2 Phase Two. During phase two, users on average took less time to review the passage compared to the initial reading, which is likely attributed to memory from the first encounter. To determine if the treatment (e.g. highlighting or LLM generated highlighting groups) significantly affected review time, we considered the percent change in reading time to review time. This accounts for individual differences, such as reading speed that may skew the results. For instance, slower readers may take a longer time to read the passage and review their notes.

We conducted a left-tailed Mann-Whitney U test to assess whether the percent change in review time for Group 1 was significantly smaller than that of Group 2. The null hypothesis assumed there was no difference in review time between Group 1 and 2. We found that there was not significant evidence (**p=0.6667**) that the tool affected review time.

We also conducted a two-tailed paired t-test to evaluate whether using our tool significantly improved evaluation scores. The null hypothesis assumed there was no difference in evaluation scores between the control and test group. The results of our analysis indicated no statistically significant difference between the evaluation scores between the two groups (**p=0.45**).

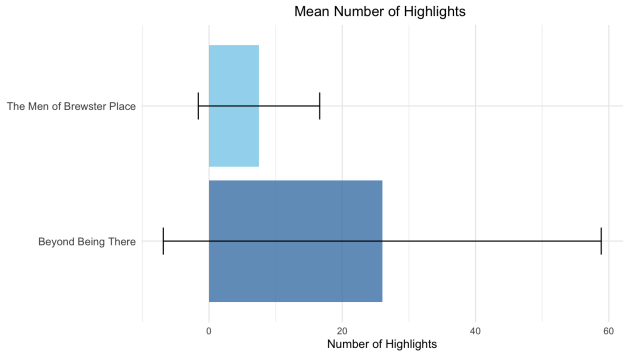

Figure 2: The average number of highlights for *Beyond Being There* was on average higher than that of *The Men of Brewster Place*. This makes sense, as it is a longer passage. However, some participants highlighted significantly more than other participants or none at all.

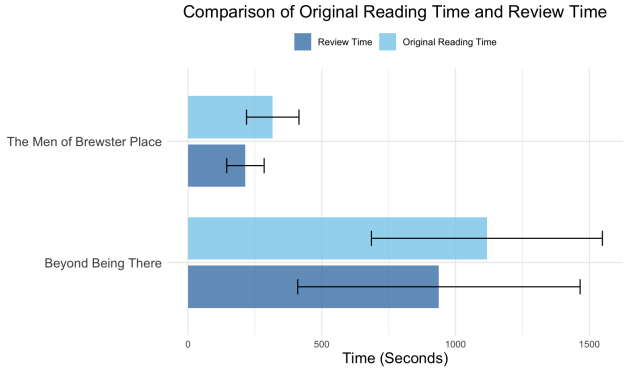

Figure 3: The average review time for Beyond Being There was higher than for The Men of Brewster Place. This aligns with expectations given the passage length. However, variability among participants suggests differences in individual reading and review strategies.

3.2 User Feedback

3.2.1 Phase One. During the initial reading, users found the tool easy to understand and use for the most part. There were minor visual bugs that were disorienting but could be dispelled by refreshing the page. As one user noted, "*The highlighting feature was a little glitchy, so fixing that would improve my experience.*" – P2

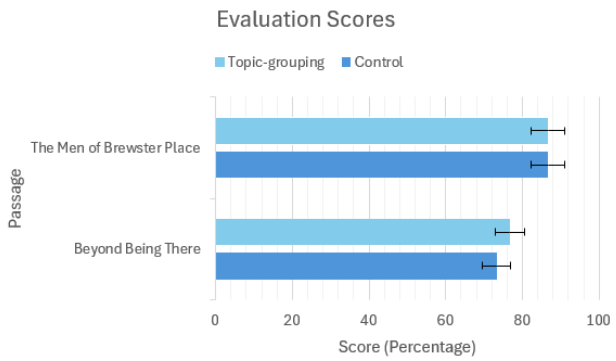

Figure 4: Scores for Beyond Being There and The Men of Brewster Place were similar across both conditions, with minor variations. This consistency implies the tool's effect on evaluation outcomes may not be substantial.

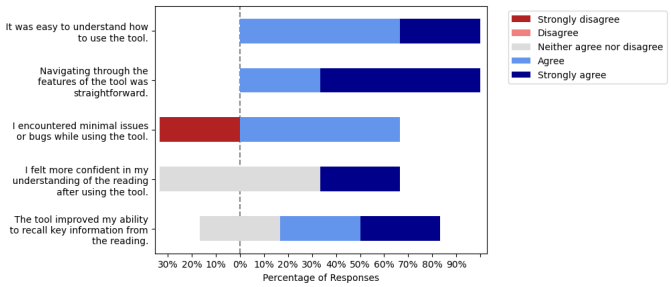

Figure 5: We asked users general Likert-scale questions about their experience with the tool. Most users found the tool intuitive, easy to understand, and confidence-boosting. However, a notable drawback was the presence of minor visual bugs, which detracted from the overall user experience. Addressing these issues through debugging could significantly improve usability.

Users would also attempt to group topics during the start of the experiment to understand the capabilities of the tool. As they read, users ceased using the group topic mechanism until a final use at the end of the passage, which may signal that users are looking to situate the accumulation of highlights that they have marked throughout the reading. Users also found the topic grouping mechanism interesting, as they stated that "*The outline corresponding to my highlights was helpful.* " – P1.

3.2.2 Phase Two. During the second phase, participants occasionally revisited the existing topics that were grouped when using the tool. In particular, for the longer passage, one participant felt encouraged to use the tool to revisit an earlier definition–"*I must return to the quote in the beginning. Let me return to what they mean by mechanism.*" – P2.

After the second phase, participants completed a post-experiment questionnaire with Likert-scale questions. Overall, participants reported enjoying the experience of using the tool. While it did not significantly improve review time or scores, participants found the tool easy to understand and helpful for recalling key information. It's encouraging that users found the tool easy to use. By applying user-centric design practices, the tool can be further enhanced to meet user needs. Participants also indicated that they would recommend it to peers for similar tasks, highlighting its potential value for others in comparable contexts.

4 Discussion

4.1 Limitations and Future Work

Our studies sample of undergraduate students is small and unrepresentative of the general American populous. Our results may not capture variations in reading across context, culture, and different age groups. We also recognize the implicit biases as researchers and how they may have affected how we designed the study and conducted the analysis. These biases, along with the nature of the feedback we received, may have influenced the prioritization of certain features or functionalities in the development of the learning tool. Future work will be directed at a more diverse group of readers with an iterative design to mitigate these limitations.

The thematic grouping tool depends on the capabilities of the GPT-4o mini model, which may introduce non-deterministic behavior. This could lead to inconsistent user experiences, especially when identical highlights produce varying thematic clusters. Addressing this issue would involve fine-tuning the model or providing users with an option to lock in desired themes.

The interval between the highlighting phase and the review phase was set at 10 days. However, this interval may not have been sufficient to fully replicate the real-world challenges of long-term retention. Future work will experiment with longer delays to better evaluate the durability if the tool's benefits over time.

The interface went through a few rounds of improvement, but usability issues still remain. Participants reported difficulties distinguishing certain highlight colors and encountered the occasional bugs in the grouping feature. One user noted that "*I think it would be helpful to also have a requirement to write a little blurb about the highlighted section.*" Enhancing interface customization option could improve user satisfaction and broaden accessibility.

Finally, highlighting is an acquired skill that can enhance knowledge retention and comprehension *when used correctly*. However, due to time constraints, we were unable to provide participants with formal training on how to highlight effectively[9]. This disparity in understanding how to highlight could have impacted the results, as participants may have used highlighting in varying ways, which could influence the outcomes in terms of comprehension and recall.

4.2 Implications for Learning

The results of our study revealed intriguing insights into how our thematic grouping tool might impact learning. While the statistical results do not confirm a significant advantage of using the highlighting tool, the trends observed in data, along with theoretical considerations, suggest that the tool may foster deeper engagement and understanding of the passages. The tool might provide a more

structured engagement with the passages. Allowing users to highlight passages and then automatically organizing the highlights into groups encourages user to focus on key content. Although the automatic grouping itself does not require critical thinking, the organization of the highlighted material into themes may help users visually categorize and connect related concepts, which can aid in understanding and remembering the information. One potential implication is that thematic grouping may facilitate the organization of information into meaningful clusters, which aligns with cognitive theories of chunking that state that people remember information better when it is grouped into coherent blocks. The process of presenting the highlights into a structured format might reduce cognitive load during review and allow participants to focus on higher-level understanding [11]. While further research is necessary to validate all of these findings, the results observed suggest that our tool could enhance learning experiences.

Users have expressed increased confidence when using the tool, even though the benefits may be limited. This heightened confidence could be a valuable metacognitive effect. Just by having an LLM validate a users preconceived notions of their highlights may suggest the application of this tool as an educational support item.

Another implication is that *there might not be a shortcut to learning*. In fact, research has shown that taking more time to deeply process information improves retention. For instance, making text harder to read, such as with obfuscated fonts, has been shown to enhance recall [4, 13]. Similarly, for spatial digital tasks, it has been found that *effort-inducing interfaces*, or interfaces that are carefully parameterized to be frustrating can actually increase engagement and spatial learning [3]. Therefore, designing a tool solely to improve review time may not align with the ultimate goal of enhancing comprehension and recall. Instead, there may be a delicate balance between creating an interface that is both engaging and challenging enough to encourage users to spend more time processing and reflecting on the information.

5 Conclusion

This study introduces and evaluates a thematic grouping tool for auto-highlighting using LLMs, aimed at improving the review process for student by organizing manually annotated highlights into coherent groups. While our quantitative results did not demonstrate significant reductions in review time or improvements in comprehension scores, the tool maintained comparable comprehension while also maintaining user satisfaction. Variability in user interaction with the tool, combined with qualitative feedback, suggest that thematic grouping may have some merit, but it remains future work to replicate the study with more participants. The results also highlight the importance of user familiarity with effective highlighting strategies, as different highlighting techniques introduced confounding variables. Overall, this study demonstrates that achieving substantial learning improvements requires a careful balance of cognitive effort and usability. Iterative refinements will allow further exploration of the tool's potential to improve long-term retention and broader applicability across diverse learning contexts.

References

[1] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047, 2019.

[2] Ed H. Chi, Lichan Hong, Michelle Gumbrecht, and Stuart K. Card. Scenthighlights: highlighting conceptually-related sentences during reading. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, page 272–274, New York, NY, USA, 2005. Association for Computing Machinery.

[3] Andy Cockburn, Per Ola Kristensson, Jason Alexander, and Shumin Zhai. Hard lessons: effort-inducing interfaces benefit spatial learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 1571–1580, New York, NY, USA, 2007. Association for Computing Machinery.

[4] Connor Diemand-Yauman, Daniel M Oppenheimer, and Erikka B Vaughan. Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, 118(1):111–115, 2011.

[5] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.

[6] Michael C Friedman. Notes on note-taking: Review of research and insights for students and instructors. *Harvard Initiative for Learning and Teaching*, pages 1–34, 2014.

[7] Jim Hollan and Scott Stornetta. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 119–125, 1992.

[8] Nikhita Joshi and Daniel Vogel. Constrained highlighting in a document reader can improve reading comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[9] Detlev Leutner, Claudia Leopold, and Viola den Elzen-Rump. Self-regulated learning with a text-highlighting strategy. *Zeitschrift für Psychologie/Journal of Psychology*, 215(3):174–182, 2007.

[10] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. The semantic reader project. *Commun. ACM*, 67(10):50–61, September 2024.

[11] George Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, 1956.

[12] Gloria Naylor. Act reading practice test questions tips, 1998.

[13] Daniel M Oppenheimer and Michael C Frank. A rose in any other font would not smell as sweet: Effects of perceptual fluency on categorization. *Cognition*, 106(3):1178–1194, 2008.

[14] Annie Piolat, Thierry Olive, and Ronald T Kellogg. Cognitive effort during note taking. *Applied cognitive psychology*, 19(3):291–312, 2005.

[15] Jessica Zeitz Self, Rebecca Zeitz, Chris North, and Alan L. Breitler. Autohighlighter: Identifying salient sentences in text. In *2013 IEEE International Conference on Intelligence and Security Informatics*, pages 260–262, 2013.

[16] Carole L Yue, Benjamin C Storm, Nate Kornell, and Elizabeth Ligon Bjork. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review*, 27:69–78, 2015.