Optimizing Human Working Memory for Text Processing and Information Retention: An AI-Enhanced Adaptive Framework for Real-Time Reading Support

Ziqian Fu, Northeastern University, fu.ziq@northeastern.edu

ABSTRACT

Many readers currently face challenges in balancing rapid text processing with deep comprehension, leading to either superficial understanding or inefficient rereading. This unstructured approach reduces study efficiency and cause readers, especially students and researchers to gradually overly reliant on AI for content summaries and analysis, filtering essays they deem worth reading, which may result in missing valuable texts and hinder deep analysis.

Unlike existing passive AI tools, this study introduces an adaptive, real-time interactive system based on Cognitive Load Theory and Active Processing Effect to support real-time reading. The system helps readers process basic text and store key ideas, thereby freeing up working memory for higher-level thinking such as analysis and synthesis.

This system offers chunked text reading mode, real-time in-place word simplifications, and sentence explanations to reduce cognitive load and prevent reading flow disruptions. Automated reflection prompts encourage active recall, and a concept diagram is generated to create a visual knowledge map, enhancing memory retention.

Delivered as a Google Chrome extension, this framework optimizes working memory, improves text processing, and boosts knowledge retention. Inspired by cognitive science theories, it provides a new way of deep reading in the digital age, fosters collaboration between technology and human cognition, freeing up cognitive resources for higher-level thinking.

CCS CONCEPTS

• Artificial Intelligence • Information Retrieval • Human-Computer Interaction

KEYWORDS

AI-assisted reading, Working Memory, Cognitive Load Theory, Active Learning, Real-time Interaction, Improvement of Reading Efficiency, Knowledge Network Construction

INTRODUCTION

According to the capacity theory of comprehension, working memory is a limited resource shared between low-level decoding and high-level understanding [1]. When basic linguistic processing demands are high, such as with unfamiliar vocabulary or complex sentence structures, less capacity remains available for deeper comprehension processes like inference and synthesis. Just and Carpenter estimate this allocation to be approximately 70% for basic processing and only 30% for analysis and synthesis [1].

This imbalance suggests inefficient working memory allocation, especially for readers with limited cognitive capacity or under cognitive load. Consequently, many readers struggle to maintain integrative understanding while decoding difficult texts during reading, since most of their cognitive resources are exhausted by basic processing tasks.

In response to this mental fatigue, readers increasingly rely on AI-powered tools to summarize and pre-analyze content. While such tools save time and effort, readers also risk undermining active engagement and deeper comprehension by outsourcing critical thinking to automated systems.

This study aims to address the root cause of passive reading behaviors —namely, the lack of cognitive reading strategies for efficient text processing and information retention. To tackle this issue, I propose an AI-enhanced real-time reading framework designed to optimize the allocation of cognitive resources. As illustrated in **Figure 1**, the system reduces the working memory load associated with basic linguistic processing—from an estimated 70% to 20%—thereby reallocating capacity toward higher-order comprehension tasks such as analysis and synthesis, which increase from 30% to 80%. Additionally, the system incorporates active retention tools, including automated reflection prompts and conceptual mapping graphs, to facilitate knowledge construction and support long-term memory retention.

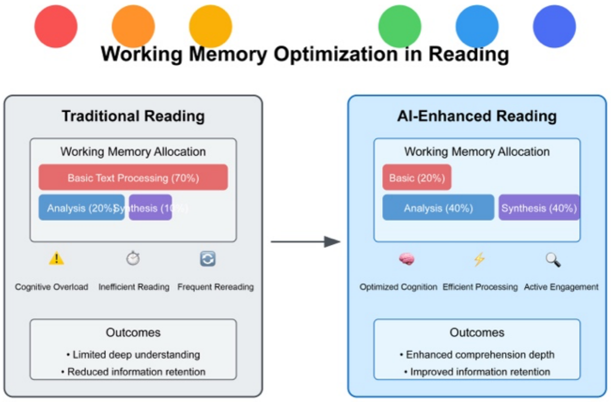

Working Memory Optimization in Reading
Traditional Reading
Working Memory Allocation
Basic Text Processing (70%)
Analysis (20%)
Synthesis (10%)
Cognitive Overload
Inefficient Reading
Frequent Rereading
Outcomes
• Limited deep understanding
• Reduced information retention
AI-Enhanced Reading
Working Memory Allocation
Basic (20%)
Analysis (40%)
Synthesis (40%)
Optimized Cognition
Efficient Processing
Active Engagement
Outcomes
• Enhanced comprehension depth
• Improved information retention

Figure 1: Cognitive resource allocation before and after AI-enhanced reading support

RELATED WORK

Working memory plays a crucial role in reading comprehension by enabling the processing and retention of information necessary for constructing coherent text representations. The study highlights that training working memory can lead to significant improvements in reading comprehension skills [2]. During reading, readers rely on the working memory to interpret vocabulary, extract key information from individual sentences, integrate ideas across paragraphs, so to grasp the overall meaning of the article.

Despite the growing popularity of AI-assisted reading tools, several limitations persist across current solutions:

Content Summary Tools (e.g., Summly): These tools offer rapid access to key points but often strip away essential context and nuance, leading to potential misunderstandings. Because the summaries are fully AI-generated, they bypass opportunities for critical thinking and user-driven interpretation.

Contextual Enhancement Tools (e.g., Scholarly): While effective in simplifying terminology and providing background information, this tool also presents automatically generated prompts that visually catch the reader's attention. Although these prompts are helpful, their presence may disrupt the reading flow and shift reader's cognitive engagement from self-reflection to AI-suggested insights, potentially reducing deep processing and active comprehension.

Intelligent Annotation Systems (e.g., Hypothesis): These platforms enable notetaking and collaboration but lack real-time cognitive scaffolding. Users must invest extra effort to manually integrate annotations, which can fragment the reading experience.

Visualization Tools (e.g., Connected Papers): Although these tools visualize thematic connections, the auto-generated quality is inconsistent, and they fail to show cross-document conceptual links, limiting their value for interdisciplinary exploration.

Interactive Research Assistants (e.g., Elicit): These systems encourage pre-reading engagement and automate literature synthesis. However, they may promote over-reliance on AI, reducing self-guided exploration. Additionally, their summaries are often generic or shallow, constrained by their underlying semantic database.

Collectively, these tools prioritize automation and efficiency, but often neglect active learning, real-time adaptive guidance in response to user needs, and reflective thinking—key components for deep comprehension and knowledge construction.

As illustrated in **Figure 2**, optimizing working memory can be achieved through multiple strategies grounded in cognitive science theories, including Cognitive Load Theory, the Active Processing Effect, the Multimedia Principle, and Distributed Cognition. These theoretical frameworks inspire my system implementations that reduce basic low-level processing load, boost the high-level analysis and synthesis load, and enhance user engagement.

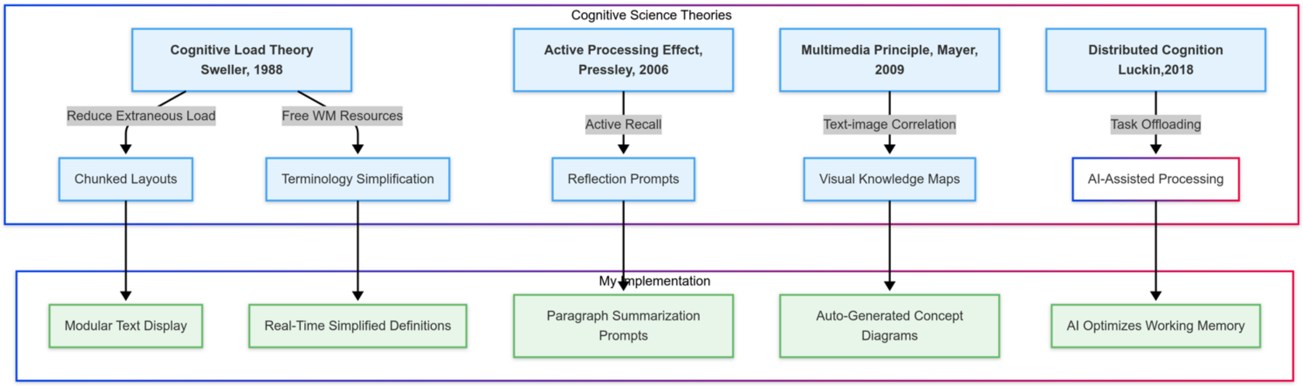

Cognitive Science Theories
Cognitive Load Theory Sweller, 1988
Active Processing Effect, Pressley, 2006
Multimedia Principle, Mayer, 2009
Distributed Cognition Luckin,2018
Reduce Extraneous Load
Free WM Resources
Active Recall
Text-image Correlation
Task Offloading
Chunked Layouts
Terminology Simplification
Reflection Prompts
Visual Knowledge Maps
AI-Assisted Processing
My Implementation
Modular Text Display
Real-Time Simplified Definitions
Paragraph Summarization Prompts
Auto-Generated Concept Diagrams
AI Optimizes Working Memory

Figure 2: Features inspired by cognitive science theories to optimize working memory

Working memory can be enhanced through strategies derived from Cognitive Load Theory [3]. Chunked Text Layout lowers intrinsic load by segmenting complex text into smaller units, allowing users to focus on one idea at a time, minimizing the cognitive burden of processing multiple ideas at the same time. This aligns with the segmenting principle of Cognitive Load Theory and improves processing efficiency by offloading low-level cognitive tasks. In addition, Terminology Simplification reduces extraneous load by removing barriers posed by unfamiliar terms, while also increasing germane load by freeing up resources for building mental models and integrating new knowledge. This informs the system's real-time, context-aware simplified definition feature. Together, these strategies optimize working memory use, foster deeper comprehension, and support schema development for more effective reading.

Inspired by the Active Processing Effect [4], the system integrates features that prompt users to actively engage with the text—such as chunk-by-chunk summarization and reflection cues. These strategies promote deeper cognitive processing, reduce passive cognitive load, and enhance working memory efficiency by reallocating mental resources toward meaningful understanding and schema development. Moreover, this active recall process supports long-term memory retention.

Guided by Mayer's Multimedia Principle [5], the system integrates visual elements alongside verbal explanations to leverage dual-channel processing in working memory. This design simplifies complex textual information by extracting key ideas and presenting them through conceptual diagrams. These diagrams make intrinsic relationships among ideas more explicit, allowing users to visually integrate information and make inferences. By distributing cognitive processing across both visual and verbal channels, the multimedia approach not only enhances working memory efficiency, but also supports deeper comprehension and long-term memory retention.

Informed by the theory of Distributed Cognition [6], the system is designed to extend the user's cognitive processes through external supports such as chunked text, pop up window, and interactive prompts. These features offload demands on working memory, reduce intrinsic and extraneous cognitive load, and provide persistent memory anchors that enhance retention. By distributing cognition across the interface and time, the system facilitates more efficient learning and deeper understanding.

My innovation lies in transforming static cognitive strategies into dynamic, user-adaptive mechanisms, bridging theory and real-time interaction:

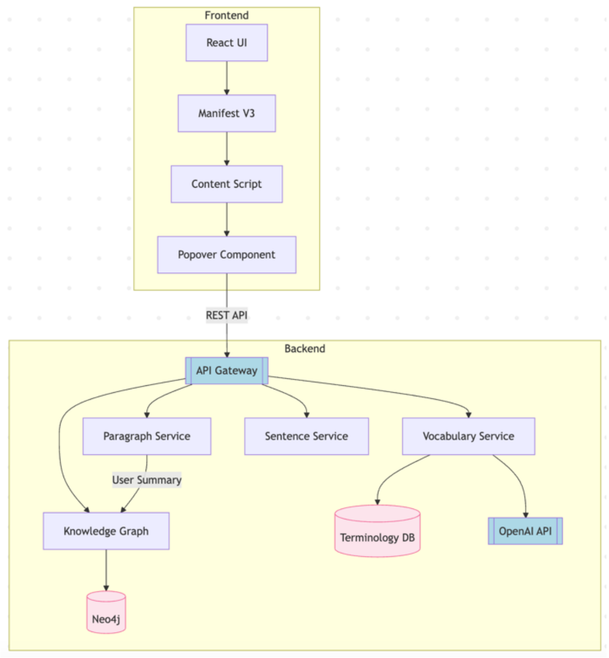
- From Cognitive Load Theory, I move beyond static content reduction toward dynamic load adaptation. Users can optionally chunk text and highlight jargon or sentences to trigger real-time simplification pop-ups—significantly freeing up working memory resources.
- Building on the Active Processing Effect, rather than offering post-reading summaries like most tools, my system integrates reflection prompts at the end of each paragraph, encouraging active recall, keeping the context information, and boosting memory retention.
- Inspired by the Multimedia Principle, instead of using fixed text-image pairs, my system provides adaptive visualizations, generated in response to users' summaries and contextual needs, enhancing conceptual integration.
- Extending Distributed Cognition, unlike fully automated and passive summarization tools, my framework promotes human-AI co-construction of summaries and syntheses, fostering deeper analytical engagement and shared cognitive processing.

SYSTEM DESIGN AND METHODOLOGY

My AI-enhanced reading framework design is based on the principles below.

- Cognitive load optimization: According to the cognitive load theory, I minimize the extrinsic cognitive load (such as complex vocabulary and syntax) while promoting the related cognitive load (such as concept understanding and information processing).
- Active Participation Facilitation: The system is designed to follow the principle of active learning, encouraging readers to actively participate in the information processing process through prompts and guidance rather than providing answers directly.
- Working Memory Enhancement: By solving basic comprehension barriers in real time, working memory capacity is released for higher-order analytical and reasoning activities.
- Knowledge integration support: Assist readers to identify and connect relevant concepts in the text and build a coherent knowledge structure.

ARCHITECTURE



The proposed framework is implemented as a Chrome extension using Manifest V3 and is supported by a modular microservices backend. The architecture is divided into two main components:

Frontend (Chrome Extension – Manifest V3)

- Implements a **chunked reading interface**, presenting one paragraph at a time to minimize cognitive load.
- Users advance through the text by pressing *Enter*, encouraging reflection before progressing.

Backend (Microservices Architecture – Node.js, REST API, and Ollama API).
The backend consists of several specialized services, each responsible for a distinct cognitive support function:

- **Vocabulary Simplification Service**: Detects highlighted terms and provides simplified synonyms or conceptual explanations with examples.
- **Sentence Understanding Service**: Analyzes selected sentences to extract the main idea, key terms, and illustrative examples.

- **Paragraph Summarization Service**: Prompts users to summarize each paragraph, reinforcing active engagement and retention.

- **Knowledge Graph Service**: Generates a visual representation of interconnected concepts based on user summaries, supporting long-term knowledge construction.

CORE FUNCTIONAL MODULES

Vocabulary Simplification Module

When a reader clicks on a word within the text, the system triggers a context-sensitive popup designed to reduce lexical complexity without interrupting reading flow. For general academic vocabulary, the popup provides a simplified synonym to enhance immediate understanding. For domain-specific technical terms, the system offers a concise concept explanation accompanied by an example of usage. Additionally, a "Simplify" button allows users to request further simplification if the initial output remains unclear, supporting incremental vocabulary acquisition and flexible learning needs.

Sentence Understanding Module

When a user highlights a sentence, the system activates a backend process that analyzes the sentence structure to extract its main idea, key terms, and a representative example. This information is displayed in a popup window designed to clarify complex or dense content. A "Simplify" button is also included, enabling users to reduce the sentence's complexity further based on their comprehension needs, thereby promoting incremental understanding and reader autonomy.

Text Chunking Module

To support focused reading and reduce cognitive overload, the system presents the text in a chunked layout—one paragraph at a time. Users proceed through the content sequentially by pressing Enter, which prompts them to reflect and generate a summary of the current paragraph before continuing. This interaction ensures deep engagement and reinforces comprehension through active recall.

Paragraph Summarization Module

At the end of each paragraph, the system poses a reflective prompt encouraging the user to summarize the key points in their own words. This method leverages the Active Processing Effect to enhance memory retention and support the development of higher order thinking skills.

Knowledge Graph Construction Module

Following each user-generated summary, the system extracts core concepts and relationships, dynamically adding a visual node to a right-hand-side knowledge graph. This growing diagram helps readers visualize how ideas interconnect across the text, reinforcing schema building and facilitating long-term knowledge integration.

User Interaction Flows

A typical user interaction flow is as follows:

- Upon visiting a webpage in Google Chrome, users can optionally activate the extension by clicking its icon. Once enabled, the system initiates chunked text mode, presenting the content paragraph by paragraph to reduce cognitive load.
- When encountering unfamiliar words, highlight and choose to get a simplified explanation.
- Faced with difficult sentences, highlight to obtain main structure and examples.
- After completing the paragraph reading, respond to the system prompts to summarize yourself.
- At the end of the reading, review the automatically generated knowledge graph to consolidate what you have learned

LESSON LEARNED

This study underscores the transformative potential of integrating cognitive science principles with adaptive computing to enhance the reading experience. The development of the AI-enhanced framework revealed that technology, when thoughtfully designed, can augment—rather than replace—human cognitive processes.

By applying the theory of Distributed Cognition, the system offloads lower-level linguistic tasks to AI, allowing readers to redirect their working memory toward higher-order comprehension tasks such as analysis, synthesis, and reflection.

A key insight from this work is the importance of achieving a delicate balance between automated assistance and user agency. Over-automation risks diminishing user engagement, while insufficient support may overwhelm readers under cognitive load. As illustrated in **Figure 3**, this proposed framework addresses this by integrating critical support functions—including text chunking, vocabulary simplification, sentence explanations, reflection prompts, and dynamic knowledge graph generation—within a seamless interface. This unified design minimizes cognitive disruption by reducing the need to switch between disparate tools, promoting sustained focus and deeper learning.

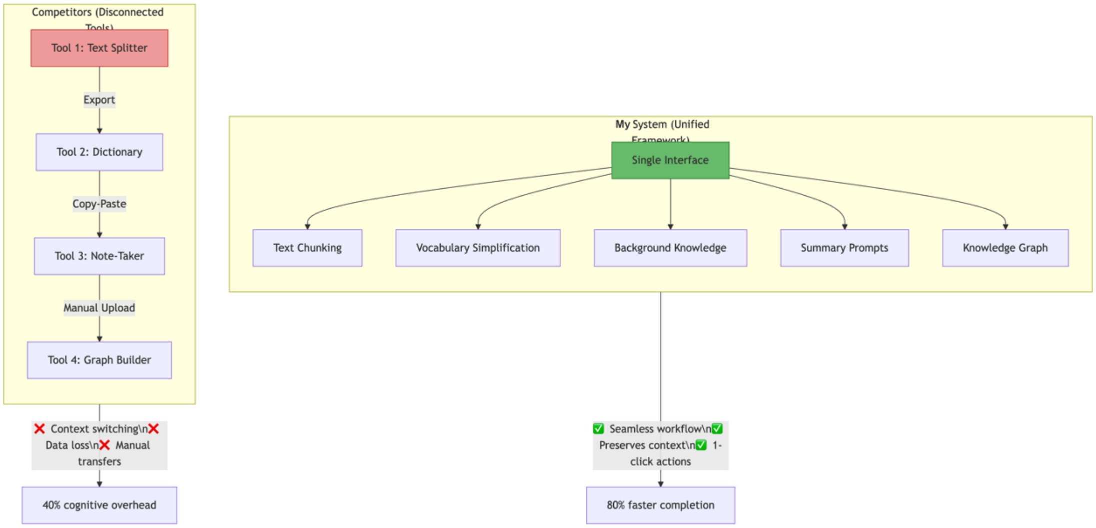


Figure 3: The competitor's disconnected tools versus this system's unified framework

FUTURE WORK

Personalized Adaptive Assistance

Future iterations can integrate real-time user modeling to tailor support dynamically based on individual reading patterns, cognitive load levels, and domain expertise. This personalization will ensure that assistance is neither excessive nor insufficient, aligning with each reader's evolving needs.

Domain-Specific Optimization

Expand the system's terminology simplification capabilities to adapt across specialized disciplines (e.g., biology, computer science) and varied textual genres (e.g., academic essays, literary narratives), ensuring accurate and context-relevant explanations.

Collaborative Reading Ecosystems

Enable multi-user interaction within the platform to support shared annotations, discussion threads, and collaborative synthesis. Such features can foster social learning and the co-construction of knowledge among readers.

Metacognitive Skill Development

Design scaffolding mechanisms that gradually reduce reliance on system prompts and instead promote the development of users' independent metacognitive reading strategies—such as self-questioning, summarization, and inference-making—empowering readers to become more self-regulated and strategic learners over time.

REFERENCES

[1] Just, M.A. and Carpenter, P.A. 1992. A capacity theory of comprehension: Individual differences in working memory. Psychological Review. 99, 1 (1992), 122–149. DOI: https://doi.org/10.1037/0033-295X.99.1.122

[2] Karin I. E. Dahlin. 2011. Effects of working memory training on reading in children with special needs. Reading and Writing, 24, 4 (April 2011), 479–491. https://doi.org/10.1007/s11145-010-9238-y

[3] Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. Cognitive Science. 12, 2 (1988), 257–285. DOI: https://doi.org/10.1207/s15516709cog1202_4

[4] Pressley, M. 2006. *Reading Instruction That Works: The Case for Balanced Teaching*. Guilford Press, New York, NY.

[5] Mayer, R.E. 2009. Multimedia Learning (2nd ed.). Cambridge University Press, New York, NY.

[6] Luckin, R. 2018. Machine Learning and Human Intelligence: The Future of Education. UCL Press, London, UK

[7] Fu, Z. 2025. AI-Enhanced Reading System – Word Munch. GitHub. Retrieved April 23, 2025 from https://github.com/iamziqian/word-munch