Beyond Linear Digital Reading: An LLM-Powered Concept Mapping Approach for Reducing Cognitive Load

Junzhi Han
Emory University
Atlanta, GA 30322
molly.han@emory.edu

Jinho D. Choi
Emory University
Atlanta, GA 30322
jinho.choi@emory.edu

Abstract

This paper presents an LLM-powered approach for generating concept maps to enhance digital reading comprehension in higher education. While particularly focused on supporting neurodivergent students with their distinct information processing patterns, this approach benefits all learners facing the cognitive challenges of digital text. We use GPT-4o-mini to extract concepts and relationships from educational texts across ten diverse disciplines using open-domain prompts without predefined categories or relation types, enabling discipline-agnostic extraction. Section-level processing achieved higher precision (83.62%) or concept extraction while paragraph-level processing demonstrated superior recall (74.51%) in identifying educationally relevant concepts. We implemented an interactive web-based visualization tool https://simplified-cognitext.streamlit.app that transforms extracted concepts into navigable concept maps. User evaluation (n=14) showed that participants experienced a 31.5% reduction in perceived cognitive load when using concept maps, despite spending more time with the visualization (22.6% increase). They also completed comprehension assessments more efficiently (14.1% faster) with comparable accuracy. This work demonstrates that LLM-based concept mapping can significantly reduce cognitive demands while supporting non-linear exploration.

1 Introduction

Complex academic texts in higher education present cognitive challenges for students who must process and retain extensive information across diverse disciplines. These challenges are particularly pronounced for neurodivergent students, including those with ADHD (Attention Deficit Hyperactivity Disorder), who process information differently and often struggle to identify and retrieve central ideas from traditional linear texts despite recognizing their importance (Ben-Yehudah and Brann, 2019; Yeari et al., 2018).

These challenges are further amplified in digital reading environments. Rather than attempting to improve traditional digital reading directly, we propose concept maps as an alternative digital interface that transforms linear text into interactive visual knowledge structures, bypassing linear reading challenges entirely while maintaining comprehensive coverage of educational content. These visual representations externalize knowledge structures, potentially reducing cognitive load while supporting the visual-spatial processing strengths often seen in students with attention-related learning differences (Sperotto, 2016; Sweller, 1988). Rather than attempting to improve traditional reading directly, concept maps provide an alternative knowledge access method that bypasses linear reading challenges entirely while maintaining comprehensive coverage of educational content.

Despite advances in automated concept mapping, significant gaps remain in current tools. Existing automated approaches typically rely on rule-based systems or predefined ontologies that lack flexibility across different domains and disciplines, often struggling with domain-specific terminology and conceptual relationships. Furthermore, existing approaches frequently extract concepts without adequately capturing the nuanced relationships between them, resulting in concept maps that lack the semantic depth necessary for comprehensive understanding.[1]

This paper investigates how large language models (LLMs) can generate comprehensive concept maps from educational texts across diverse academic disciplines. Rather than enhancing traditional digital reading, we propose concept maps as an alternative educational interface that trans-

[1]All our resources including source codes and concept map data are publicly accessible through our open source project: https://github.com/emorynlp/cognitext

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 805–817
July 31 - August 1, 2025

forms linear text into interactive visual knowledge structures, enabling students to access the same information through non-linear exploration. We examine three research questions:

1. How effectively can LLMs identify key concepts across diverse academic disciplines without domain-specific training or ontologies?

2. What differences exist in the extraction and representation of knowledge relationships across different academic disciplines?

3. To what extent do automatically generated concept maps reduce cognitive load and improve reading comprehension compared to traditional linear reading?

Our approach implements concept extraction to identify key terms and ideas, relation identification to determine semantic connections between concepts, and concept map generation to organize these elements into visual knowledge structures. We evaluate system performance across ten academic disciplines and assess the impact of resulting concept maps on reading comprehension and cognitive load.

This work makes several key contributions: (1) we establish a methodological framework for LLM-based concept extraction across different types of academic content; (2) we provide empirical evidence for domain-specific knowledge structures that inform adaptive concept mapping; (3) we demonstrate the practical application of language models for educational concept map generation; and (4) we present evidence that concept map visualization as a reading alternative significantly reduces cognitive load while maintaining or improving comprehension outcomes.

While each system component (preprocessing granularity, LLM-based extraction, and user evaluation) merits individual investigation, this work demonstrates their integration for practical educational applications. We focus on the educational impact as our primary contribution, with detailed component analysis providing supporting evidence for system design decisions.

2 Related Work

2.1 Cognitive Load in Educational Contexts

Cognitive Load Theory, developed by (Sweller, 1988), distinguishes between three types of mental processing: intrinsic load (inherent task complexity), extraneous load (poor instructional design), and germane load (meaningful learning processes). Educational interventions that reduce extraneous load while maintaining or enhancing germane load can greatly improve learning outcomes, particularly for students with diverse cognitive processing patterns.

Complex academic texts impose substantial cognitive demands through dense information presentation, abstract concept relationships, and linear narrative structures that may not align with individual learning preferences. For neurodivergent students, these challenges are particularly pronounced. Le Cunff et al. (2024) demonstrated that neurodivergent students, particularly those with ADHD, reported significantly higher extraneous cognitive load compared to neurotypical peers, while showing no differences in intrinsic or germane cognitive load. This pattern suggests that the presentation format of educational materials, rather than their inherent complexity, creates disproportionate challenges.

While digital reading environments present additional complications—with neurobiological research by Zivan et al. (2023) revealing higher cognitive load patterns in screen-based reading—the fundamental challenge extends beyond medium to the linear, text-heavy presentation of complex information. Bahari et al. (2023) identified several approaches that successfully manage cognitive load in educational environments, including visualization-based approaches and argument mapping, with most strategies aimed at reducing extraneous cognitive load while fostering germane load through generative learning practices.

Concept maps specifically address cognitive load by transforming extraneous processing demands into germane learning opportunities. By externalizing knowledge structures and providing visual-spatial representations, concept maps can reduce the working memory burden of maintaining conceptual relationships while reading, allowing students to focus cognitive resources on understanding and connecting ideas rather than tracking linear narrative flow.

2.2 Automated Concept and Relation Extraction for Education

Automated concept extraction and concept map generation has evolved from rule-based approaches to sophisticated machine learning methods. Early computational approaches established foundations for extracting concept maps from educational texts, with Aguiar et al. (2018) providing comprehensive approaches for concept maps mining from text.

Recent work has applied large language models to concept map generation. Perin et al. (2023) demonstrated automated concept map generation using fine-tuned large language models, while other work has shown that LLMs can identify conceptually complex regions of text (Garbacea et al., 2021) and extract concepts from academic materials while preserving semantic relationships (Zhang et al., 2023).

Educational relation extraction differs from general-purpose approaches in its emphasis on pedagogically meaningful connections that support learning progression (Dessì et al., 2020). Recent advances in prompt-based approaches have shown particular promise for educational contexts. Chen et al. (2022) introduced KnowPrompt, incorporating knowledge from relation labels into prompt construction. Advancing this, Chen et al. (2024) developed a Generative context-Aware Prompt-tuning method (GAP) that eliminates the need for domain experts to design prompts. Large language models can extract relational knowledge when prompted appropriately (Jiang et al., 2020), with models like GPT-3.5 and GPT-4 demonstrating competitive performance in processing domain-specific educational content with minimal training requirements (Hu et al., 2024).

The key challenge lies in distinguishing concepts and relations of varying pedagogical importance while capturing both local conceptual connections and broader structural relationships that reflect disciplinary knowledge organization. Our work builds on these foundations by implementing hierarchical concept classification and multi-level relation identification for cross-disciplinary educational applications.

2.3 Visualization Techniques for Knowledge Representation

Concept maps, originally developed by Novak (Novak, 1998) and refined by Novak and Cañas (Novak and Cañas, 2008), organize knowledge through explicit propositions with labeled relationships (e.g., "Photosynthesis → produces → Oxygen"). Unlike mind maps (Buzan, 1993), which support ideational writing through brainstorming via radial structures, concept maps employ hierarchical representations optimized for reading comprehension and systematic knowledge acquisition.

Meta-analyses demonstrate educational effectiveness: Anastasiou et al. (2024) found a moderate positive effect of concept maps on science achievement ($g = 0.776$) across 55 studies, while Schroeder et al. (2018) reported similar benefits of learning with concept maps on educational outcomes ($g = 0.58$) across 142 studies. Recent neurological research by Shealy et al. (2022) demonstrated that concept mapping alters cognitive activation patterns, increasing activity in brain regions associated with divergent thinking. Yang et al. (2025) addressed cognitive load concerns by introducing progressive concept maps that integrate information incrementally, improving learning outcomes compared to conventional approaches.

Our approach builds on Novakian concept mapping theory specifically for reading support rather than writing facilitation. Alternative text representation frameworks include Schema Theory (Bartlett, 1932; Rumelhart, 1977), van Dijk and Kintsch's macrostructures (van Dijk and Kintsch, 1983), and Rhetorical Structure Theory (Mann and Thompson, 1988). Our work combines LLM contextual understanding with comprehensive knowledge representation across diverse disciplines, implementing progressive disclosure to manage cognitive load in educational concept maps.

3 Methodology

We developed a systematic approach for extracting concepts and relationships from educational texts and transforming them into interactive concept maps. Figure 1 illustrates the key components of our methodology.

3.1 Dataset Selection and Preparation

We selected ten Wikipedia articles representing diverse academic disciplines: biology, mathematics/statistics, computer science, linguistics, art, history, philosophy, political science, health/medicine, and one general non-academic field. Articles were chosen based on specific criteria to represent content that undergraduate students would likely encounter during their academic studies but would not be familiar with from prior education.

Each selected article maintained sufficient conceptual depth, a neutral academic tone, and introduced new concepts rather than common basics. The articles ranged from 1,383 to 11,337 words (mean: 4,839). Preprocessing steps included

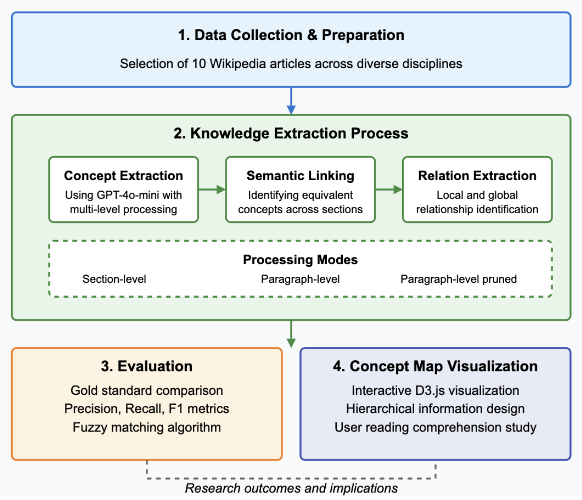

Figure 1: Overview of the methodology for concept & relation extraction, and concept map visualization.

HTML removal, section identification, reference removal, and tokenization using spaCy.

3.2 Text Processing Modes

We implemented three distinct text processing modes to evaluate how different granularities of input text affect extraction quality:

1. **Section-level processing** utilizes complete sections from articles as input units, enabling the system to process larger chunks of coherent text and potentially capture broader thematic relationships. Each complete section was provided as single input to GPT-4o-mini for concept and relation extraction.

2. **Paragraph-level processing** operates on individual paragraphs, allowing the system to focus on local concepts and relationships within more concentrated contexts. Individual paragraphs were processed separately as independent input units, with results from all paragraphs subsequently aggregated to form the complete concept map for each article.

3. **Paragraph-pruned processing** applies post-processing filters to the results obtained from paragraph-level processing to address noise and irrelevant concepts. After completing the standard paragraph-level extraction described above, we applied two filtering mechanisms: first, elimination of concepts appearing exclusively in single paragraphs to reduce noise from isolated mentions; second, semantic filtering using the allMiniLM-L6-v2 transformer model by Reimers and Gurevych (2019) to calculate similarity scores between concepts and their section contents, preserving only concepts with similarity scores above 0.6.

3.3 Extraction Framework

3.3.1 Concept Extraction

The concept extraction process utilizes GPT-4o-mini with precise prompting strategies to identify educationally relevant concepts within academic texts. A concept is defined as "a significant term or phrase that represents a fundamental idea, entity, or phenomenon within a discipline."

The extraction methodology implements a hierarchical schema that categorizes concepts into three distinct layers based on educational significance:

1. **Priority Layer (Core Concepts)**: Fundamental principles, key terminology, major themes, and critical processes (15-20%)

2. **Secondary Layer (Supporting Concepts)**: Sub-processes, related theories, and component parts (40-50%)

3. **Tertiary Layer (Contextual Elements)**: Author contributions, specific examples, and historical developments (30-40%)

We developed a specialized prompt structure that explicitly targets concepts answering fundamental knowledge questions ("what," "how," "why," and "when"), ensuring comprehensive coverage across knowledge dimensions.

3.3.2 Relation Extraction

The relation extraction framework defines semantic connections between previously extracted concepts as structured triplets consisting of a source concept, a target concept, and a descriptive relation type. The extraction employs a multi-tiered approach:

1. **Local Relations**: Connections between concepts within the same textual segment (section or paragraph)

2. **Global Relations**: Higher-order connections between concepts across different sections

This dual-layer approach addresses the challenge of aligning relation extraction with learning objectives by prioritizing pedagogically meaningful connections. Local relations establish foundational concept understanding within focused contexts, while global relations reveal broader conceptual frameworks essential for comprehensive domain knowledge. Combined with our hierarchical schema (Priority, Secondary, Tertiary layers) from concept extraction (Section 3.3.1), this ensures that extracted relations support progressive learning objectives, with core concepts and their relationships receiving priority in visualization to align with pedagogical goals of foundational understanding before detailed exploration. Each identified relationship includes supporting evidence from the source text that validates its authenticity and enhances educational value, ensuring extracted relations support structured learning goals rather than arbitrary semantic associations.

3.4 Evaluation Framework

To establish a gold standard dataset, two undergraduate annotators independently performed complete manual concept and relation extraction from each article. The first annotator had greater familiarity with STEM disciplines, while the second had stronger background in liberal arts and humanities. Each annotator identified all educationally relevant concepts and their relationships within each text, creating comprehensive manual annotations that served as ground truth for evaluating our automated extraction performance.

Concepts were rated on a 4-point scale (0-3) assessing educational significance from irrelevant (0) to core concepts (3). Relations were evaluated on a 3-point scale (0-2) based on pedagogical utility. Inter-annotator agreement was assessed using Cohen's Kappa coefficient, with values of $\kappa = 0.76$ for concepts and $\kappa = 0.71$ for relations, indicating substantial agreement between annotators despite their different disciplinary backgrounds.

The evaluation employed precision, recall, and F1 scores, comparing our automated extractions against the manually created gold standard annotations. Fuzzy matching was applied to accommodate linguistic variability between automated and manual annotations. The matching algorithm applied a hierarchical process beginning with exact string comparisons and progressively applying more flexible techniques based on edit distance and word-level similarity to identify semantically equivalent concepts and relations across the two annotation sets.

We acknowledge that using undergraduate annotators rather than domain experts across all ten disciplines represents a limitation in our evaluation methodology, as disciplinary expertise could affect the identification of field-specific concepts and relationships.

3.5 Concept Map Visualization

We implemented an interactive concept map visualization system called Cognitext using D3.js force-directed layouts. The interface, shown in Figure 2, incorporates several features designed to support effective navigation:

1. **Hierarchical information architecture**: Priority concepts are displayed with the darkest node coloring to indicate their fundamental importance, while secondary and tertiary concepts remain initially hidden to prevent cognitive overload

2. **Self-directed exploration**: When priority concepts have connections to hidden lower-level concepts, they display a pulsing orange indicator circle that signals available deeper exploration and encourages user interaction

3. **Visual focus management**: Selecting concepts brings related nodes to the foreground while fading others

4. **Relationship transparency**: Hovering over connections reveals relationship types and supporting textual evidence

5. **Intelligent content enhancement**: An integrated concept chatbot provides contextual explanations and answers questions based on the extracted knowledge

Users interact with the system through multiple methods: exploring from central nodes outward, clicking for concept explanations, and rearranging nodes to customize their view. The implementation uses Python for backend processing and Streamlit Cloud for web deployment: https://simplified-cognitext.streamlit.app.

3.6 User Evaluation

To assess the efficacy of concept maps for reading comprehension, we conducted a controlled study with 14 undergraduate participants (8 female, 6 male; ages 19-24; 4 self-identified as ADHD). While the sample size is modest, it provides initial insights into concept map effectiveness.

Concept Map Generation. For the user study, concept maps were generated using section-level processing results, which achieved the highest average F1 score in our evaluation. To ensure valid

Discipline	Section-Level P	Section-Level R	Section-Level F1	Paragraph-Level P	Paragraph-Level R	Paragraph-Level F1	Paragraph-Pruned P	Paragraph-Pruned R	Paragraph-Pruned F1
CS	81.53	59.40	68.72	55.71	69.41	61.81	64.89	67.54	66.19
Biology	89.86	68.17	77.52	59.92	79.62	68.38	69.10	76.75	72.72
History	85.47	65.83	74.38	62.29	81.35	70.55	71.43	75.09	73.21
Philosophy	80.63	53.70	64.46	56.42	72.13	63.31	65.51	70.25	67.80
Politics	83.92	67.15	74.61	61.93	81.63	70.43	68.19	72.76	70.40
Linguistics	82.14	49.62	61.87	50.92	62.62	56.17	60.12	63.75	61.88
Art	83.21	63.54	72.06	58.32	75.02	65.62	66.46	70.15	68.25
Math	79.13	58.37	67.18	53.94	71.62	61.53	63.18	68.83	65.88
Medicine	82.98	66.82	74.03	54.66	73.28	62.61	67.84	69.49	68.65
General	87.35	69.18	77.21	60.77	78.38	68.46	71.95	74.60	73.25
Average	83.62	62.18	71.20	57.49	74.51	64.89	66.87	70.92	68.82

Table 1: Performance of concept extraction by discipline with fuzzy matching (P/R/F1 = Precision/Recall/F1 Score).

assessment of effectiveness, the automatically extracted concepts and relations were manually corrected using our gold standard annotations before visualization. This correction process removed incorrectly identified concepts, added missing educationally relevant concepts, and refined relationship labels for accuracy and clarity.

Experimental Design. Participants read two academic articles, linguistics and physics, with comparable college-level complexity (Flesch-Kincaid scores: 16.9 and 17.2). Articles were selected based on complexity metrics rather than participant domain expertise. One article was read using traditional linear reading and the other through the concept map interface. Article assignment and reading sequence were counterbalanced across participants to mitigate order effects.

Assessment Methodology. Following each reading task, participants completed a comprehensive assessment designed to evaluate understanding across different conceptual levels supported by the hierarchical concept map structure. The assessment included: 10 factual recall questions on priority concepts (multiple choice), 5 conceptual understanding questions requiring integration of priority and secondary level knowledge (short answer), 3 relationship identification tasks, and 2 knowledge transfer problems requiring synthesis of information from multiple concept layers. Questions were developed by the lead researcher based exclusively on the original source texts, without reference to extracted concept maps or their structural limitations.

Measurements. Cognitive load was assessed using the mental demand dimension of the NASA Task Load Index (Hart and Staveland, 1988), a validated multidimensional rating scale. While the full NASA-TLX measures perceived workload across six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration), we focused specifically on the mental demand subscale as it most directly relates to cognitive load in reading comprehension tasks. Participants rated mental demand on a 1-10 scale, with scores ranging from 1 (very low mental demand) to 10 (very high mental demand). Additionally, semi-structured interviews were conducted to gather qualitative feedback on user experience and perceived effectiveness of the concept mapping approach.

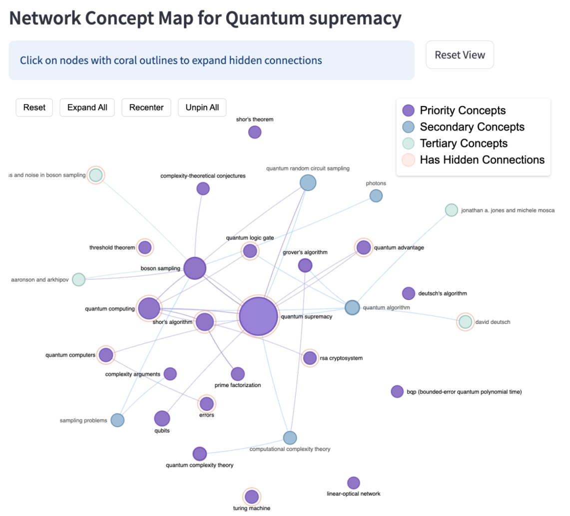

Figure 2: Cognitext interface showing the interactive concept map with labeled relationships between concepts (visible on hover), exploration features, and concept chatbot.

4 Results

4.1 Extraction Performance

Table 1 presents the performance metrics for concept extraction across various academic disciplines using fuzzy matching. We compared three distinct text processing approaches.

Section-level processing demonstrated superior precision across all disciplines, achieving an average precision of 83.62%, with biology (89.86%) and general domain (87.35%) articles showing highest performance. However, this approach showed comparatively lower recall (62.18% average), indicating that while it extracted highly relevant concepts, it missed some concepts present in the gold standard dataset.

Paragraph-level processing yielded considerably higher recall metrics (74.51% average), with political science (81.63%) and history (81.35%) articles showing highest recall. This improvement came at the cost of precision, which dropped to 57.49% average. The paragraph-level pruned approach achieved intermediate performance in both precision (66.87%) and recall (70.92%).

When comparing F1 scores, section-level processing performed best overall (71.20% average), followed by paragraph-level pruned (68.82%) and standard paragraph-level (64.89%) approaches.

Relation extraction performance followed similar patterns across processing approaches and disciplines (detailed results in Appendix A Table 3). Section-level processing achieved highest overall precision (78.61%), with biology and general articles showing particularly strong performance (82.09% and 83.51% respectively) and moderate recall (59.76%). Paragraph-level processing exhibited substantially lower precision (51.95%) but higher recall (69.08%), while the pruned approach demonstrated intermediate performance with 62.01% precision and 67.29% recall. Overall, section-level processing performed best with an average F1 score of 67.71%, followed by paragraph-level pruned (64.52%) and standard paragraph-level (59.28%) approaches.

4.2 Concept/Relation Distribution Patterns

Figure 3 presents a heatmap visualization of the normalized concept distribution across academic disciplines. When normalized for text length, the philosophy article showed the highest concept density (31.09 concepts per 1,000 words), followed by health/medicine (30.93) and political science (27.66). In contrast, the history (9.61) and art (10.86) articles exhibited the lowest density.

Normalization revealed discipline-specific patterns in concept distribution. The computer science article emphasized problems & solutions and mathematical foundations (both 3.14 concepts per 1,000 words). The health/medicine article showed a pronounced focus on medical & safety concepts (7.73) and processes & mechanisms (6.87). The philosophy article demonstrated a great emphasis on core concepts (6.51) and socio-cultural contexts (4.34).

For relation distribution, Table 5 in Appendix A shows that structural relations emerged as the most frequent relation type across all disciplines, with the health/medicine (12.89), philosophy (11.57), and political science (10.19) articles showing the highest normalized frequencies. The overall relation density varied substantially across disciplines, with health/medicine exhibiting the highest density (54.22 relations per 1,000 words), followed by philosophy (45.55) and political science (41.48).

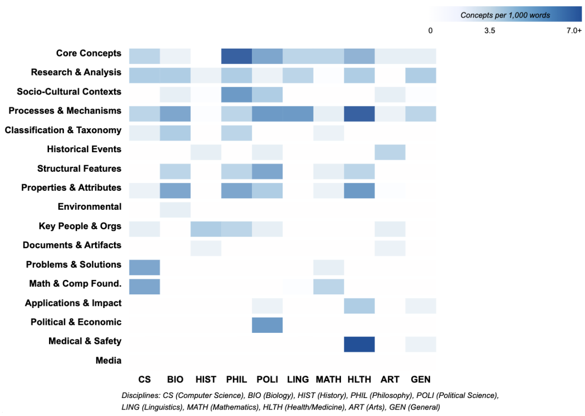

Figure 3: Concept distribution heatmap across academic disciplines, normalized per 1,000 words. Color intensity represents frequency, darker blue = higher density.

4.3 User Study Results

To evaluate concept map effectiveness on reading comprehension, we compared traditional linear reading with concept map-assisted reading (Table 2). Participants spent more time with the concept mapping tool (32.5 vs. 26.5 minutes, 22.6% increase) but completed comprehension assessments more quickly (18.3 vs. 21.3 minutes, 14.1% decrease).

Most significantly, participants reported substantially reduced perceived mental effort when using concept maps (5.0 vs. 7.3 on NASA TLX scale), representing a 31.5% decrease in cognitive load that was statistically significant (p = 0.00092). Comprehension accuracy showed slight improvement (98% vs. 97%), though not statistically significant.

We acknowledge that our assessment instrument focused on propositional knowledge, which concept maps represent effectively, and this should be considered when interpreting the comparable accuracy results.

User feedback was generally positive (mean rating 4.21/5), with participants highlighting the tool's ability to "visually analyze basic concepts" and "show hierarchical relationships between different concepts."

Metric	Without Tool	With Tool
Reading Time	26.5 min	32.5 min
Assessment Time	21.3 min	18.3 min
Mental Effort	7.3	5.0
Correctness	97%	98%

Table 2: Comparison of reading performance with and without concept mapping tool. Reading time refers to text reading (linear condition) vs. concept map exploration (concept map condition)

5 Analysis and Discussion

5.1 Extraction Performance and Disciplinary Patterns

The consistent precision-recall trade-off across processing approaches demonstrates a fundamental tension in concept extraction methodology. Section-level processing achieved higher precision through comprehensive contextual understanding, while paragraph-level processing captured more concepts through granular analysis. The substantial precision advantage for relations (78.61% vs. 51.95%) suggests that relational semantics often depend on broader contextual understanding than can be captured within paragraph boundaries, aligning with discourse coherence theory that semantic relationships emerge from macro-level textual structures.

The paragraph-level pruned approach effectively mitigated many limitations of standard paragraph-level processing while preserving much of its recall advantage. This suggests that incorporating multi-stage validation processes into extraction pipelines can substantially improve performance without requiring contextual windows as large as section-level processing.

Cross-disciplinary analysis revealed systematic knowledge organization patterns reflecting epistemological differences between fields. Scientific text (biology) showed superior extraction metrics across all approaches, suggesting more explicit externalization of conceptual relationships through standardized linguistic patterns. Historical text performed better with paragraph-level processing, indicating conceptual relationships are established within localized narrative units rather than extended theoretical frameworks. The linguistics article's extraction challenges highlight complexities in meta-disciplinary discourse where language is both medium and subject of analysis.

The distinctive patterns in concept and relation distribution reflect disciplinary epistemologies. Philosophy's high density of causal relations but absence of functional relations suggests emphasis on conceptual reasoning, while health/medicine's high functional relation density points to procedural knowledge focus. These findings suggest that universal knowledge representation approaches may not be optimal, and concept maps might benefit from tailoring to specific relation structures observed in different academic content types.

5.2 Concept Map Effectiveness for Cognitive Load Reduction

The comparison of concept mapping visualization to traditional linear reading reveals a complex relationship between time investment, cognitive load, and comprehension outcomes. The observed increase in reading time (22.6%) paired with a decrease in assessment time (14.1%) when using the concept mapping tool suggests a shift in cognitive resource allocation. While users invested more time in initial exploration, they subsequently completed assessment tasks more efficiently.

The substantial reduction in perceived mental effort (31.5%) despite longer engagement time represents one of our most significant findings. This relationship suggests the visualization transformed extraneous cognitive load into germane cognitive load, enabling more productive mental processing rather than simply reducing overall demands. This transformation is particularly valuable for educational applications where sustained engagement with complex material is desirable.

While the marginal improvement in comprehension accuracy (1%) appears modest, this should be interpreted within the context of the already high baseline performance (97%), suggesting a potential ceiling effect. The combination of comparable comprehension outcomes with significantly reduced cognitive effort indicates an improved efficiency ratio—participants achieved similar results with less mental strain.

The generally positive user feedback (4.21/5) confirms participants recognized value in the visualization approach. The implementation of hierarchical information architecture with progressive disclosure aligns with cognitive load theories by preventing information overload while maintaining access to comprehensive content, enabling incremental mental model construction while preserving underlying concept connections.

5.3 Implications for Educational Technology

These findings suggest that domain-agnostic extraction systems face inherent limitations, and maximizing extraction performance might benefit from adaptive approaches tailored to different discourse types. The cognitive load reduction without compromising comprehension indicates concept mapping tools could be particularly valuable for students experiencing cognitive fatigue during traditional reading, including those with attention-related difficulties.

Morever, the implementation of hierarchical information architecture with progressive disclosure aligns with cognitive load theories by preventing information overload while maintaining access to comprehensive content. The observed self-directed exploration through concept maps aligns with constructivist learning principles, suggesting these tools can support diverse learning approaches and accommodate individual differences in background knowledge and processing styles while transforming extraneous cognitive load into germane cognitive load.

6 Conclusion and Future Work

This work demonstrates that LLM-based concept mapping can significantly reduce cognitive demands while supporting non-linear exploration of educational content. Our findings suggest concept mapping tools particularly benefit students experiencing cognitive fatigue, transforming linear text into interactive visualizations that support self-directed exploration.

Future work should prioritize three key directions based on our empirical findings:

Adaptive Extraction Methodologies. Our analysis revealed systematic variation across disciplines, with philosophy articles showing highest concept density (31.09 per 1,000 words) while history articles exhibited lowest (9.61). Future research should develop adaptive methodologies that automatically adjust to disciplinary discourse patterns by: (1) implementing discourse pattern recognition to select optimal processing granularity, (2) developing domain-specific relation taxonomies based on our finding that structural relations dominate in health/medicine (12.89 per 1,000 words) while causal relations are prominent in philosophy (10.85 per 1,000 words), (3) creating hierarchical processing pipelines that combine section-level precision (83.62%) with paragraph-level recall (74.51%), and (4) developing automated quality assurance methodologies that can refine extracted concepts and relations without manual intervention.

Longitudinal Impact Studies. While our study (n=14) demonstrated 31.5% cognitive load reduction, critical questions remain about long-term educational impact. Future studies should examine knowledge retention and transfer beyond immediate comprehension, conduct targeted research with larger samples of neurodivergent students (our study included only four self-identified participants), and investigate whether concept maps sustainably reduce cognitive load across time and contexts.

LMS Integration. Current implementation barriers limit practical deployment. Priority efforts should focus on developing plugins for major learning management systems (Canvas, Blackboard, Moodle), implementing collaborative features for shared editing and instructor annotations, and addressing processing latency through scalable architectures for institutional deployment.

These directions address our core finding that concept mapping transforms extraneous cognitive load into germane cognitive load while maintaining comprehension outcomes.

Limitations

Despite the promising results of this research, several limitations should be acknowledged when interpreting our findings.

Corpus limitations. Our analysis was restricted to examining only one article per academic discipline, which limits generalizability. The findings should be interpreted as article-specific observations that suggest potential disciplinary patterns rather than definitive characterizations of entire domains. Future work should expand the corpus to include multiple texts from each discipline to establish more generalizable patterns.

Ecological validity limitations. Our use of Wikipedia articles, while providing standardized, neutral content across disciplines, may not fully represent typical educational materials such as textbooks, journal articles, or course-specific readings. Wikipedia's encyclopedic structure and hyperlinked format may facilitate concept extraction differently than traditional academic texts. Future work should evaluate performance on authentic course materials to establish broader applicability.

Evaluation methodology limitations. Our evaluation relied on undergraduate annotators rather than domain experts, and comprehension questions were developed by the lead researcher rather than assessment professionals. While inter-annotator agreement was substantial and questions assessed multiple understanding levels, expert involvement would strengthen future evaluations.

Generalizability across learning differences. While we hypothesized particular benefits for students with attention-related learning differences, our sample included only a small number of self-identified neurodivergent participants (n=4). More targeted research with larger samples of neurodivergent students would be necessary to substantiate claims about differential benefits across students.

Assessment methodology limitations. The node-and-edge structure of concept maps excels at representing straightforward propositions (e.g., "Photosynthesis produces Oxygen") but struggles with conditional relationships (e.g., "A causes B only under specific conditions"), complex temporal sequences, or counterfactual reasoning. Consequently, our question development may have inadvertently excluded assessment items requiring these more complex cognitive operations, potentially favoring the concept map condition. While this limitation does not invalidate our findings within the scope of propositional knowledge comprehension, it restricts generalizability to the full spectrum of reading comprehension skills typically assessed in educational contexts.

Technical constraints. The use of GPT-4o-mini, while cost-efficient, introduced model-specific constraints including knowledge cutoff limitations, reduced parameter capacity compared to larger models, and context window restrictions that particularly affected global relation extraction. These limitations may have impacted extraction performance, especially for longer documents. The system also demonstrated significant processing latency with longer articles, which could limit practical deployment in time-sensitive educational contexts.

Implementation challenges. The current implementation faces practical deployment barriers including limited integration with existing learning management systems, basic visualization capabilities without advanced features like collaborative editing, and minimal customization options for educators. These constraints, while providing clear directions for future development, limit immediate broad adoption in educational settings.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback that significantly improved this work. We are grateful to our user study participants for their time and valuable insights. Special thanks to the Computer Science Department of Emory University and the Emory NLP Lab for providing computational resources and technical support. We are grateful to everyone who contributed to the creation of Cognitext and supported our efforts to improve educational comprehension for all learners.

References

Camila Zacche Aguiar, Davidson Cury, and Amal Zouaq. 2018. Towards technological approaches for concept maps mining from text. *CLEI Electronic Journal*, 21(1):7.

Dimitris Anastasiou, Clare Nangsin Wirngo, and Pantelis Bagos. 2024. The effectiveness of concept maps on students' achievement in science: A meta-analysis. *Educational Psychology Review*, 36(39):1–18.

Akbar Bahari, Sumei Wu, and Paul Ayres. 2023. Improving computer-assisted language learning through the lens of cognitive load. *Educational Psychology Review*, 35:53.

Frederic C. Bartlett. 1932. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Gal Ben-Yehudah and Adi Brann. 2019. Pay attention to digital text: The impact of the media on text comprehension and self-monitoring in higher-education students with ADHD. *Research in Developmental Disabilities*, 89:120–129.

Tony Buzan. 1993. *The Mind Map Book: Unlock Your Creativity, Boost Your Memory, Change Your Life*. BBC Books, London.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, pages 1–11, New York, NY, USA. ACM.

Zhenbin Chen, Zhixin Li, Yufei Zeng, Canlong Zhang, and Huifang Ma. 2024. GAP: A novel generative context-aware prompt-tuning method for relation extraction. *Expert Systems with Applications*, 248:123478.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2020. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *CoRR*, abs/2011.01103.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097. Association for Computational Linguistics.

Sandra G Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, pages 139–183. North-Holland.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Anne-Laure Le Cunff, Vincent Giampietro, and Eleanor Dommett. 2024. Neurodiversity positively predicts perceived extraneous load in online learning: A quantitative research study. *Education Sciences*, 14(5).

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Joseph D. Novak. 1998. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Lawrence Erlbaum Associates.

Joseph D. Novak and Alberto J. Cañas. 2008. The theory underlying concept maps and how to construct and use them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition.

Wagner Perin, Davidsom Cury, Camila Aguiar, and Crediné Menezes. 2023. From text to maps: Automated concept map generation using fine-tuned large language model. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1317–1328, Porto Alegre, RS, Brasil. SBC.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David E. Rumelhart. 1977. Schemata: The building blocks of cognition. In Rand J. Spiro, Bertram C. Bruce, and William F. Brewer, editors, *Theoretical Issues in Reading Comprehension*, pages 33–58. Lawrence Erlbaum Associates.

N.L. Schroeder, J.C. Nesbit, C.J. Anguiano, and 1 others. 2018. Studying and constructing concept maps: a meta-analysis. *Educational Psychology Review*, 30:431–455.

T. Shealy, J. S. Gero, and P. Ignacio. 2022. How the use of concept maps changes students' minds and brains. In *ASEE Annual Conference and Exposition, Conference Proceedings*.

L. Sperotto. 2016. The visual support for adults with moderate learning and communication disabilities: How visual aids support learning. *International Journal of Disability, Development and Education*, 63(2):260–263.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press.

Kai-Hsiang Yang, Hui-Chun Chu, Gwo-Jen Hwang, and Tzu-Jung Liu. 2025. A progressive concept map-based digital gaming approach for mathematics courses. *Educational Technology Research and Development*.

M. Yeari, E. Vakil, L. Schifer, and R. Schiff. 2018. The origin of the centrality deficit in individuals with attention-deficit/hyperactivity disorder. *Journal of Clinical and Experimental Neuropsychology*, 41(1):69–86.

Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-based interactive exploration and customization of document summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Michal Zivan, Sasson Vaknin, Nimrod Peleg, Rakefet Ackerman, and Tzipi Horowitz-Kraus. 2023. Higher theta-beta ratio during screen-based vs. printed paper is related to lower attention in children: An EEG study. *Plos one*, 18(5):e0283863.

A Appendix

A.1 Detailed Extraction Performance

Table 3 presents the detailed performance metrics for relation extraction across the ten academic disciplines. Similar to concept extraction, we compared three distinct text processing approaches: section-level, paragraph-level, and paragraph-level pruned processing.

	Section			**Paragraph**			**Para-Pruned**		
Disc.	P	R	F1	P	R	F1	P	R	F1
CS	79.3	56.9	66.3	49.5	63.9	55.8	60.2	64.9	62.5
Bio	82.1	66.8	73.7	56.8	74.9	64.6	65.9	74.3	69.9
Hist	78.3	66.1	71.7	57.6	76.5	65.7	66.2	70.4	68.2
Phil	79.2	50.4	61.6	50.6	66.2	57.4	59.0	67.6	63.0
Pol	76.5	62.7	68.9	52.0	75.6	61.6	61.9	68.9	65.2
Ling	78.8	47.6	59.4	46.1	57.4	51.2	55.8	59.2	57.5
Art	77.6	59.0	67.1	52.8	68.3	59.6	60.4	66.2	63.2
Math	73.4	57.7	64.7	47.9	66.2	55.6	59.7	63.2	61.4
Med	77.4	64.9	70.6	49.2	68.3	57.2	63.6	67.7	65.5
Gen	83.5	65.5	73.4	56.9	73.6	64.2	67.4	70.5	68.9
Avg	78.6	59.8	67.7	52.0	69.1	59.3	62.0	67.3	64.5

Table 3: Performance of relation extraction by discipline with fuzzy matching (P = Precision, R = Recall, F1 = F1 Score).

A.2 Concept and Relation Categorization

Table 4 presents the normalized distribution of concept types across disciplines (per 1,000 words), highlighting discipline-specific knowledge organization patterns.

Concept	**CS**	**BIO**	**HIST**	**PHIL**	**POL**	**LING**	**ART**	**MATH**	**MED**	**GEN**
Core	2.2	1.2	0.6	6.5	4.7	2.6	1.6	2.2	3.4	1.7
Research	2.7	2.6	1.1	2.9	1.1	2.2	0.0	0.9	2.6	2.5
SocioCult	0.0	1.5	0.9	4.3	2.9	0.4	1.4	0.2	0.0	1.0
Process	2.5	3.8	1.0	2.2	4.4	4.8	1.2	1.9	6.9	2.5
Classif	1.7	2.6	0.0	2.2	0.4	0.0	0.0	1.4	0.0	0.0
Struct	0.0	2.3	0.0	2.2	3.6	0.0	0.0	1.7	2.2	0.0
Property	1.2	3.2	0.6	3.6	2.9	0.0	0.9	1.0	4.7	0.8
Environ	0.0	1.5	0.6	0.0	0.0	0.0	0.0	0.3	0.0	0.0
People	1.7	0.3	2.4	2.2	1.8	0.0	1.4	0.7	0.0	0.0
Docs	0.0	0.0	1.3	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Problems	3.1	0.0	0.0	0.0	0.4	0.0	0.0	1.9	0.0	0.6
Math/Comp	3.1	0.0	0.0	0.0	0.0	0.9	0.0	2.1	0.0	0.0
Impact	0.7	0.0	0.1	0.0	1.1	0.0	0.5	0.7	3.0	1.3
Politics	0.0	0.0	0.1	0.7	4.0	0.0	0.3	0.0	0.0	0.0
Medical	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	7.7	1.3
Media	0.3	0.0	0.1	0.0	0.0	0.0	0.5	0.3	0.0	0.0
Events	0.0	0.6	1.5	0.0	1.5	0.0	2.1	0.0	0.4	0.0

Table 4: Normalized concept distribution (per 1,000 words) across all disciplines.

Table 5 presents the normalized distribution of relation types across disciplines (per 1,000 words).

A.3 Example Prompt Templates

To facilitate reproducibility, we provide examples of the prompt templates used for concept and relation extraction:

Relation Type	**BIO**	**LING**	**PHIL**	**HLTH**	**CS**
Structural	6.74	4.40	11.57	12.89	6.95
Causal	2.34	4.40	10.85	10.31	5.96
Impact	4.98	3.96	10.12	9.88	6.29
Functional	6.45	1.32	0.00	9.02	5.79
Interaction	2.05	3.52	6.51	6.01	4.14
Cognitive	0.00	3.08	0.00	0.86	0.50
Linguistic	0.00	2.20	0.00	0.43	0.17

Table 5: Normalized relation distribution (relations per 1,000 words) for selected relation types across five representative disciplines.

A.3.1 Concept Extraction Prompt (Abbreviated)

A concept is defined as a significant term or phrase that represents a fundamental idea, entity, or phenomenon within a discipline. Extract key concepts from the provided text using the following guidelines.

Concept Layers: 1. **Core Concepts (Priority Layer):** - Primary theoretical concepts and fundamental principles - Key terminology and definitions essential to the topic - Major themes and overarching frameworks

2. **Supporting Concepts (Secondary Layer):** - Sub-processes and variations of core concepts - Related theories and complementary ideas - Component parts and organizational structures

3. **Contextual Elements (Tertiary Layer):** - Author names and their key contributions - Specific examples and case studies - Historical context and developments

Output Format: [List of JSON objects with entity, context, evidence, and layer fields]

A.3.2 Relation Extraction Prompt (Abbreviated)

Extract key relationships between these available concepts using the following guidelines. The extracted relations will be used for visualizations to aid educational comprehension.

Guidelines: - Ensure that the relations are clearly defined and relevant to the text's main ideas. - Focus on capturing a variety of relationship types without restricting to specific categories. - Avoid speculative relationships; only include those with explicit or strong implicit textual support.

Output Format: [List of JSON objects with source, relation_type, target, and evidence fields]