Distribution Agreement

Junzhi Han March 27, 2025

Reducing Cognitive Load in Digital Reading: An LLM-Powered Approach for Universal Reading Comprehension

By

Junzhi Han

Jinho D. Choi
Advisor

Quantitative Theory and Method

Jinho D. Choi
Advisor

Alex Grizzell
Committee Member

Gordon Berman
Committee Member

Emily Wall
Committee Member

2025

Reducing Cognitive Load in Digital Reading: An LLM-Powered Approach for Universal Reading Comprehension

By

Junzhi Han

Jinho D. Choi
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Method

2025

Abstract

Reducing Cognitive Load in Digital Reading: An LLM-Powered Approach for Universal Reading Comprehension
By Junzhi Han

This research investigates how pre-trained large language models (LLMs) can generate concept maps to enhance digital reading comprehension in higher education. While particularly focused on supporting neurodivergent students with their distinct information processing patterns[49, 57], this approach benefits all learners facing the cognitive challenges of digital text. The study employs GPT-4o-mini to extract concepts and relationships from educational texts across ten diverse disciplines using open-domain prompts without predefined concept categories or relation types, enabling truly discipline-agnostic extraction applicable to all educational domains. Evaluation of three text processing approaches against a manually annotated gold dataset reveals that for concept extraction, section-level processing achieves the highest precision (83.62%), while paragraph-level processing demonstrates superior recall (74.51%). For relation extraction, similar patterns emerge with section-level processing showing the highest precision (78.61%) and paragraph-level processing yielding better recall (69.08%). Disciplinary variations are observed in both extraction tasks, with biology showing the strongest concept (F1=77.52%) and relation (F1=73.65%) extraction performance while humanities disciplines have comparatively lower performance. An interactive web-based visualization tool was developed that transforms extracted concepts into navigable concept maps using D3.js force-directed layouts, accessible at https://simplified-cognitext.streamlit.app/. User evaluation (n=14) revealed that while participants spent more time engaging with concept maps (22.6% increase), they experienced substantially reduced cognitive load (31.5% decrease in perceived mental effort) and completed comprehension assessments more efficiently (14.1% faster) with marginal improvements in accuracy. Qualitative feedback (mean rating: 4.21/5) highlighted the tool's effectiveness in visualizing conceptual relationships, though initial adaptation challenges were noted. This work contributes to educational technology by establishing a framework for LLM-based concept extraction, providing evidence on processing granularity effects, developing a concept categorization system for educational mapping, and creating a visualization tool with demonstrated learning benefits.

Reducing Cognitive Load in Digital Reading: An LLM-Powered Approach for Universal Reading Comprehension

By

Junzhi Han

Jinho D. Choi
Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Method

2025

Acknowledgments

As I complete this honors thesis, I wish to express my sincere gratitude to those who have supported me throughout this academic journey.

First and foremost, I would like to thank my thesis advisor, Professor Jinho Choi, for his invaluable guidance, expertise, and patience. His thoughtful feedback and encouragement were instrumental in helping me navigate the challenges of this research project. My appreciation extends to my committee members—Professor Gordon Berman, Professor Alex Grizzell, and Professor Emily Wall—whose insights significantly improved this work.

I thank the Department of Quantitative Theory and Methods and Emory University for providing the necessary resources, and the Emory NLP lab for offering valuable collaborative opportunities with experienced upperclassmen.

To my classmates and friends who offered academic support and emotional encouragement, and to my parents whose unwavering belief has been my constant strength: thank you. Special thanks to my partner, for his extraordinary dedication to this project, particularly his assistance with manual annotation and survey distribution, without which this research would not have been possible.

This academic milestone represents not just the culmination of research but also personal growth. Through this process, I have learned to approach challenges with greater calm and perspective, understanding that the pursuit of knowledge is a continuous journey rather than a destination.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital reading presents significant cognitive challenges, especially in higher education where students must process and retain extensive information from digital texts [8]. Research shows this creates barriers for all students but particularly affects those with attention-related learning differences, making tools that enhance digital comprehension increasingly important as educational institutions adopt more digital materials [5, 9, 31].

1.1 The challenge of digital reading and concept maps as a potential solution

The digital reading environment presents unique obstacles to comprehension and retention [33, 58]. Students with ADHD face particular difficulties in this environment, as they process information differently and often struggle to identify and retrieve central ideas from digital texts despite recognizing their importance [8, 49, 57]. These challenges point to the need for structured support mechanisms that can aid comprehension for diverse learners [8, 49].

Non-linear reading approaches, particularly concept maps, offer a promising solution by enabling students to navigate information according to conceptual relationships rather than predetermined sequences [1, 51]. 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 [11, 49, 50].

1.2 Current approaches and limitations

Text processing techniques can transform unstructured content into structured information, with educational applications focusing on organizing content and supporting comprehension [10, 18, 42, 60]. While specialized frameworks exist for educational content [35], significant gaps remain in current concept mapping tools.

Despite these advances, significant gaps remain in current concept mapping tools for educational content. While automated approaches exist, they typically rely on rule-based systems or predefined ontologies that lack flexibility across different domains and disciplines [15]. These tools often struggle with domain-specific terminology and conceptual relationships that vary significantly between fields such as humanities, social sciences, and STEM disciplines [39]. 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 [54].

The automation of concept and relation extraction presents substantial challenges across diverse educational domains. Educational texts vary considerably in structure, terminology density, and relationship complexity depending on the discipline. Science texts often contain explicit technical terms and clearly delineated relationships, while humanities texts feature more implicit concepts and nuanced connections [30]. Additionally, educational materials employ domain-specific relationships (such as ”is a prerequisite for” in mathematics or ”is evidence for” in scientific arguments) that require specialized extraction approaches [38, 25].

1.3 Research questions and hypothesis

This research addresses these challenges by investigating the fundamental question: “Can concept maps be generated using Large Language Models for articles from all educational domains?” To systematically explore this question, I examine the following specific sub-questions:

- **How effectively can LLMs identify key concepts across diverse academic disciplines without domain-specific training or ontologies?** This question examines the domain-agnostic extraction capabilities of large language models when applied to educational texts from humanities, social sciences, and STEM fields.

- **What differences exist in the extraction and representation of knowledge relationships across different academic disciplines?** This question explores how disciplinary discourse patterns affect both the types of relationships extracted and the organization of conceptual knowledge.

- **To what extent do automatically generated concept maps reduce cognitive load and improve reading comprehension compared to traditional linear reading?** This question investigates the practical educational benefits of transforming linear text into visual knowledge structures, particularly for students with attention-related learning differences.

This investigation follows a systematic three-step process: First, concept extraction identifies the key terms and ideas present in educational texts. Second, relation identification determines the semantic connections between these concepts. Finally, concept map generation organizes these elements into a coherent visual representation that supports comprehension.

The study is guided by two primary hypotheses:

- By using large language models for concept and relation extraction, the system is expected to generate comprehensive concept maps that maintain consistent performance across diverse educational domains including humanities, social sciences, and STEM fields.

- By transforming linear educational text into visual concept maps through automated extraction, the system is expected to reduce cognitive load and improve reading comprehension for college students, particularly benefiting those with attention-related learning differences who demonstrate strengths in visual-spatial processing.

Chapter 2

Background

2.1 Cognitive load in digital reading

Cognitive load refers to the total amount of mental effort used in working memory during learning or information-processing tasks [50]. This concept provides a foundational framework for understanding the mental demands of learning, distinguishing between three types of cognitive load: intrinsic (inherent to the task complexity), extraneous (imposed by instructional design), and germane (related to schema construction). Digital reading environments present unique cognitive challenges that differ from traditional print reading, stemming from the interaction between human cognitive architecture and digital interface characteristics. While cognitive load theory traditionally suggests minimizing extraneous load, Skulmowski and Rey [46] highlight how digital learning environments have challenged these assumptions—elements such as interactivity and immersion may introduce extraneous load while simultaneously promoting motivation and learning. This apparent contradiction indicates that cognitive load in digital environments requires a nuanced understanding that considers both positive and negative effects of design elements and aligns cognitive demands with desired learning outcomes.

Neurobiological research provides compelling evidence for the distinct cognitive demands of digital reading compared to print reading. Zivan et al. examined brain activation differences between screen and print reading in children using electroencephalogram measurements, revealing significant differences in spectral power patterns. Print reading was associated with higher energy in high-frequency bands (beta, gamma), while screen reading showed higher power in lower frequency bands (alpha, theta) [61]. The higher theta-to-beta ratio observed during screen reading indicates challenges in attention allocation, suggesting greater cognitive load and reduced focused attention during digital reading. These neurobiological findings correlated with behavioral measures of attention, providing direct physiological evidence of the different cognitive demands imposed by digital versus print formats.

These biological differences are amplified by the structural characteristics of digital texts. Hypertexts, with their non-linear organization and embedded links, present specific cognitive challenges beyond those found in traditional linear formats. Taky Eddine [52] highlights how the intricate complexities of hypertexts can lead to cognitive overload. The nearly limitless amount of immediately accessible reading resources in digital environments, coupled with complex navigation structures, creates novel reading challenges absent in traditional linear texts. Koc-Januchta et al. [24] found that perceived non-optimal design diverts cognitive resources away from meaningful processing, resulting in lower learning gains despite potentially engaging features. These findings emphasize that while digital tools can enhance learning when properly aligned with cognitive principles, poor design can undermine effectiveness by imposing unnecessary cognitive load.

Despite these neurobiological and structural challenges, research on actual performance differences between digital and print reading shows mixed results. Ocal et al. [53] found no significant differences in college students' reading comprehension between paper and screen conditions. However, students reported preferring paper-based reading for complex material, demonstrating awareness of the increased cognitive demands of digital reading for challenging content. This discrepancy between performance and preference highlights the complex interaction between medium, content complexity, and individual perceptions of cognitive effort. Bahari [6], in reviewing computer-assisted language learning strategies, identified several approaches that successfully manage cognitive load in digital environments, including online annotations, captioning, visualization-based approaches, and argument mapping. Most strategies aimed to reduce extraneous cognitive load, though some fostered germane load through generative learning practices—suggesting that effective digital reading environments can be designed when cognitive principles guide development.

These cognitive load challenges inform the design of support tools like concept maps, which may reduce the cognitive demands of digital reading by transforming linear text into visual knowledge structures. Such visual representations could be particularly valuable for students with attention-related learning differences, potentially compensating for the reduced focused attention observed during digital reading while supporting personalized interaction with content.

2.2 Reading challenges for neurodivergent students

Research on reading comprehension in students with ADHD has revealed consistent patterns of difficulty with certain aspects of comprehension. Friedman et al. conducted a scoping review of thirty-four studies examining the relationship between ADHD and reading comprehension ability. Their analysis found that reading comprehension is generally impaired in individuals with ADHD, with particularly pronounced difficulties in tasks requiring students to retell or identify central ideas in stories. However, performance varied based on task demands, with some studies showing improved performance when reading comprehension task requirements were reduced [41]. These findings suggest that students with ADHD may struggle not necessarily with basic comprehension but with the executive function demands of organizing, prioritizing, and extracting central ideas from text.

The cognitive demands of digital learning environments present particular challenges for neurodivergent students. Le et al. examined the factors impacting cognitive load in online learning for neurodivergent versus neurotypical university students. Their qualitative comparison found that while both groups reported similar challenges such as navigation difficulties and technical issues, neurodivergent students experienced additional barriers including problems with inaccurate transcripts and inaccessible content presentation [27]. These findings shows that neurodivergent and neurotypical students face similar challenges but with differing degrees of intensity.

More specifically, Petrovskaya et al. investigated the relationship between neurodiversity and perceived cognitive load in online learning. Their survey of 231 students found that neurodivergent students, particularly those with ADHD, reported significantly higher extraneous cognitive load compared to neurotypical peers, while showing no significant differences in intrinsic or germane cognitive load [26]. This pattern suggests that the design of digital learning environments, rather than the inherent complexity of the material, creates disproportionate challenges for neurodivergent students.

While the research discussed above highlights the particular challenges faced by neurodivergent students, it is important to recognize that reading difficulties are not unique to neurodivergent populations. Students who are non-native English speakers, those with limited background knowledge in a subject area, and even neurotypical students in unfavorable reading conditions can experience similar challenges with comprehension, information extraction, and cognitive load.

This recognition aligns with the principles of Universal Design for Learning (UDL), which emphasizes creating educational materials and tools that benefit all students rather than specialized interventions for specific populations. This inclusive approach acknowledges that reading comprehension challenges exist on a spectrum across diverse student populations, and that tools designed with accessibility considerations can enhance learning experiences universally, regardless of neurological profile or language background.

To transform linear text into the concept maps that could address these cognitive and attentional challenges, automated extraction of key concepts from educational text is essential.

The evolution of natural language processing techniques for concept extraction provides a foundation for understanding how large language models might generate accurate and educationally valuable concept representations across diverse disciplines.

2.3 Natural language processing for concept extraction

Concept extraction, the process of identifying key terms and ideas from text, forms the essential first step in transforming linear educational content into structured knowledge representations. Unlike traditional named entity recognition (NER), which primarily focuses on predefined categories such as people, organizations, and locations, educational concept extraction requires identifying domain-specific terminology and abstract ideas that constitute core knowledge components.

Traditional approaches to concept extraction relied primarily on statistical methods and keyword frequency analysis. These evolved into more sophisticated rule-based systems that incorporated linguistic patterns and domain-specific ontologies. While effective within narrow domains, these approaches lacked the flexibility required for cross-disciplinary educational applications, where terminology and knowledge structures vary significantly between fields.

Machine learning approaches marked a significant advancement in educational concept extraction. Chau et al. [10] developed FACE, a supervised feature-based machine learning method specifically designed for extracting concepts from digital textbooks to support adaptive navigation and content recommendation. This approach enabled more flexible identification of domain concepts but still required substantial training data. To address multi-domain challenges, Penggao [19] introduced fuzzy semantic classification for e-learning concepts, which not only extracted concepts but also semantically clustered and classified them using fuzzy membership values, supporting adaptive learning paths across domains. Feature-rich hybrid approaches further enhanced extraction capabilities, as demonstrated by Lee et al. [28], who showed that combining traditional machine learning with handcrafted linguistic features and transformer-based approaches significantly improved performance on educational content.

Large language models (LLMs) represent the current frontier in concept extraction, offering unprecedented flexibility and contextual understanding. These models can identify concepts across diverse academic disciplines without requiring domain-specific rules or extensive labeled data. The contextual awareness of LLMs enables them to distinguish between domain-specific usages of terms that might appear identical to traditional extraction methods. Garbacea et al. [20] demonstrated that such models can even identify conceptually complex regions of text that require simplification, providing valuable guidance for prioritizing concept extraction in educationally challenging content. Zhang et al. [59] showcased the practical application of these capabilities in ConceptEVA, a system that extracts concepts from research papers and visualizes them in a force-directed layout preserving both semantic relationships and co-occurrence patterns. Their "focus-on" feature allows users to select concepts of interest, bringing them to the forefront while maintaining semantic relationships with other concepts—demonstrating the interactive potential of modern concept extraction approaches.

The present research builds upon this progression, addressing limitations of existing methods by leveraging LLMs with educational domain awareness to extract concepts across diverse disciplines. This approach aims to combine the contextual understanding of LLMs [28] with the comprehensive knowledge representation goals of earlier approaches [10], while ensuring consistent performance across humanities, social sciences, and STEM fields [19]. By incorporating complexity awareness [20], the system specifically targets support for students with varying learning needs through appropriate visualization of conceptually complex elements.

2.4 Relation extraction approaches

Relation extraction (RE), the task of identifying semantic relationships between entities in text, has evolved through several methodological approaches, each offering distinct advantages for educational applications. This progression from traditional methods to contemporary approaches has particular relevance for generating concept maps that accurately represent knowledge structures across diverse academic disciplines.

Traditional rule-based approaches relied on predefined patterns and linguistic templates to identify relationships between entities. While these methods achieved high precision for well-defined relation types, they required extensive manual engineering and lacked flexibility across domains. Educational materials present unique challenges for such approaches due to their domain-specific terminology and relationship structures. General relation extraction typically focuses on predefined relationship types between named entities [42], but educational applications require identifying specialized relationships aligned with learning objectives [18] that vary significantly across disciplines [30].

Graph-based neural models represented a significant advancement by capturing both local and document-level relationships. Christopoulou et al. [14] developed an edge-oriented graph neural model creating document-level graphs with various node and edge types, enabling learning of both intra- and inter-sentence relations. This capability is essential for educational content where key concepts build upon one another across document sections. The enrichment of these models with entity-specific information, as demonstrated by Soares et al. [55], further enhanced relation classification by providing contextual understanding of educational terms—critical for accurately representing their relationships in concept maps.

Prompt-tuning approaches have emerged as particularly effective for educational relation extraction by offering flexibility and domain adaptability. Chen et al. [12] introduced KnowPrompt, incorporating knowledge from relation labels into prompt construction—an approach adaptable to educational settings by encoding domain-specific relationship types like "is a prerequisite for" or "provides evidence for." Advancing this methodology, Chen et al. [13] developed a Generative context-Aware Prompt-tuning method (GAP) that eliminates the need for domain experts to design prompt templates, incorporating a prompt generator that extracts relation triggers from context. Son et al. [47] extended these capabilities to conversational contexts through GRASP (Guiding model with RelAtional Semantics using Prompt), capturing semantic relationships in dialogue—valuable for processing educational content from discussion forums or interactive learning materials.

Large language models (LLMs) represent the current state-of-the-art in relation extraction, offering unprecedented flexibility without requiring domain-specific rules or extensive labeled data. Jiang et al. [23] demonstrated that LLMs can effectively extract relational knowledge when prompted appropriately, with automatically generated diverse prompts significantly improving extraction accuracy. This approach is particularly relevant for educational content where relationships may be expressed through varied linguistic constructions across different disciplines. Antaki et al. [22] further showed that models like GPT-3.5 and GPT-4 can process complex domain-specific data and extract meaningful relationships with minimal training data, addressing the limitations of current educational concept mapping tools that "typically rely on rule-based systems or predefined ontologies that lack flexibility across different domains and disciplines" [15].

The present research builds upon these methodological advances by combining LLM-based extraction capabilities with principles from both prompt-tuning and document-level relation extraction. This integrated approach addresses the limitations of rigid rule-based systems [15] while employing contextually-aware extraction similar to GAP [13] to adapt to the specific terminology and relationship patterns found in diverse educational domains.

2.5 Visualization techniques for knowledge representation

Knowledge representation through visualization offers powerful tools for organizing and communicating complex information structures, with concept mapping standing as one of the most well-established approaches [32, 29].

Concept maps represent a structured approach to visualizing knowledge through graphical representations of concepts and their relationships. As defined by Novak, concept maps are graphical tools for organizing and representing relationships between concepts indicated by connecting lines, with linking words or phrases that specify the nature of relationships [40]. These maps typically organize information hierarchically, from general to specific concepts, and are constructed around a central focus question that provides context and purpose to the knowledge organization. Concept maps differ from related approaches such as mind maps (which are less formal and typically radiate from a central concept) and knowledge graphs (which often employ more complex ontological structures and formal relationship types).

The educational value of concept mapping has been substantiated through rigorous meta-analyses. Anastasiou et al. [2] analyzed 55 studies involving 5,364 students in Grades 3-12, finding a moderate positive effect size for concept mapping on science achievement (g = 0.776). Their research revealed varying effects across disciplines—moderate for biology and chemistry but large for physics and earth science—suggesting domain-specific benefits. In a broader examination spanning multiple knowledge domains, Schroeder et al. [44] synthesized 142 independent effect sizes (n = 11,814) and found that learning with concept maps produced a moderate, statistically significant positive effect (g = 0.58). Notably, actively creating concept maps (g = 0.72) was more beneficial than merely studying pre-made maps (g = 0.43), highlighting the importance of engagement in the mapping process.

The structure of concept maps significantly influences their effectiveness for different knowledge representation tasks. Safayeni et al. [43] established an important theoretical distinction between traditional hierarchical concept maps and “Cyclic Concept Maps.” While traditional maps excel at representing static knowledge structures, cyclic maps are designed specifically for representing functional or dynamic relationships between concepts. This distinction highlights that different knowledge structures may require different visualization approaches. Network-based concept maps offer yet another alternative, providing flexibility for representing complex interconnections between concepts without enforcing rigid hierarchical relationships. This approach is particularly valuable for complex domains where relationships between concepts are multidirectional and non-hierarchical.

Recent research has revealed the neurological effects of concept mapping on learning processes. Shealy et al. [45] investigated how concept mapping affects engineering students’ approach to design problems, finding that students who developed concept maps produced more diverse problem statements with less semantically similar words. Neurological measurements showed that concept mapping altered cognitive activation patterns, with reduced activity in the left prefrontal cortex (associated with convergent thinking) and increased activation in the right prefrontal cortex (associated with divergent thinking). These findings provide compelling evidence that concept mapping can fundamentally alter cognitive approaches to complex problems, potentially enhancing creative problem-solving abilities.

The effectiveness of concept maps varies based on both content complexity and implementation strategy. Yang et al. [56] noted that traditional concept mapping can increase cognitive load when learning content is extensive or complex. Their research introduced a progressive concept map-based approach that integrates concepts incrementally, significantly improving learning achievement, motivation, problem-solving tendencies, and self-efficacy compared to conventional implementations. Beyond educational applications, concept mapping principles have been applied to large-scale knowledge representation systems such as ConceptNet [48], demonstrating how these approaches can be scaled to represent broad domains of knowledge.

The present research implements a network concept map approach for representing knowledge extracted from educational texts. This decision was informed by the need for flexibility in representing complex interconnections between concepts across diverse academic disciplines, allowing for a more authentic representation of the complex conceptual landscapes that characterize different fields of study.

Chapter 3

Approach

3.1 Overview of methodology

This section presents a systematic methodology for assessing the ability of large-scale language models for concept and relationship extraction in educational texts. The main goal is to assess the effectiveness of these models in identifying and extracting concepts and their relationships in order to construct educational concept maps that can be used as valuable aids for interdisciplinary learning. Figure 3.1 illustrates the key components of the methodology that will be detailed in the subsequent sections.

The methodology contains four main components. First, a gold standard dataset was created by a manual annotation process through rigorous manual evaluation. The process involved two independent annotators who examined articles from ten different disciplines against detailed conceptual and relational annotation guidelines. Inter-annotator agreement was assessed using Cohen's Kappa coefficient to ensure annotation reliability.

Second, the automated extraction process utilizes GPT-4-mini to perform concept and relation extraction at different granularities of text processing. Extraction operations are performed at both chapter and paragraph levels, using an incremental approach that processes text blocks sequentially. The method first extracts initial concepts from the first two text

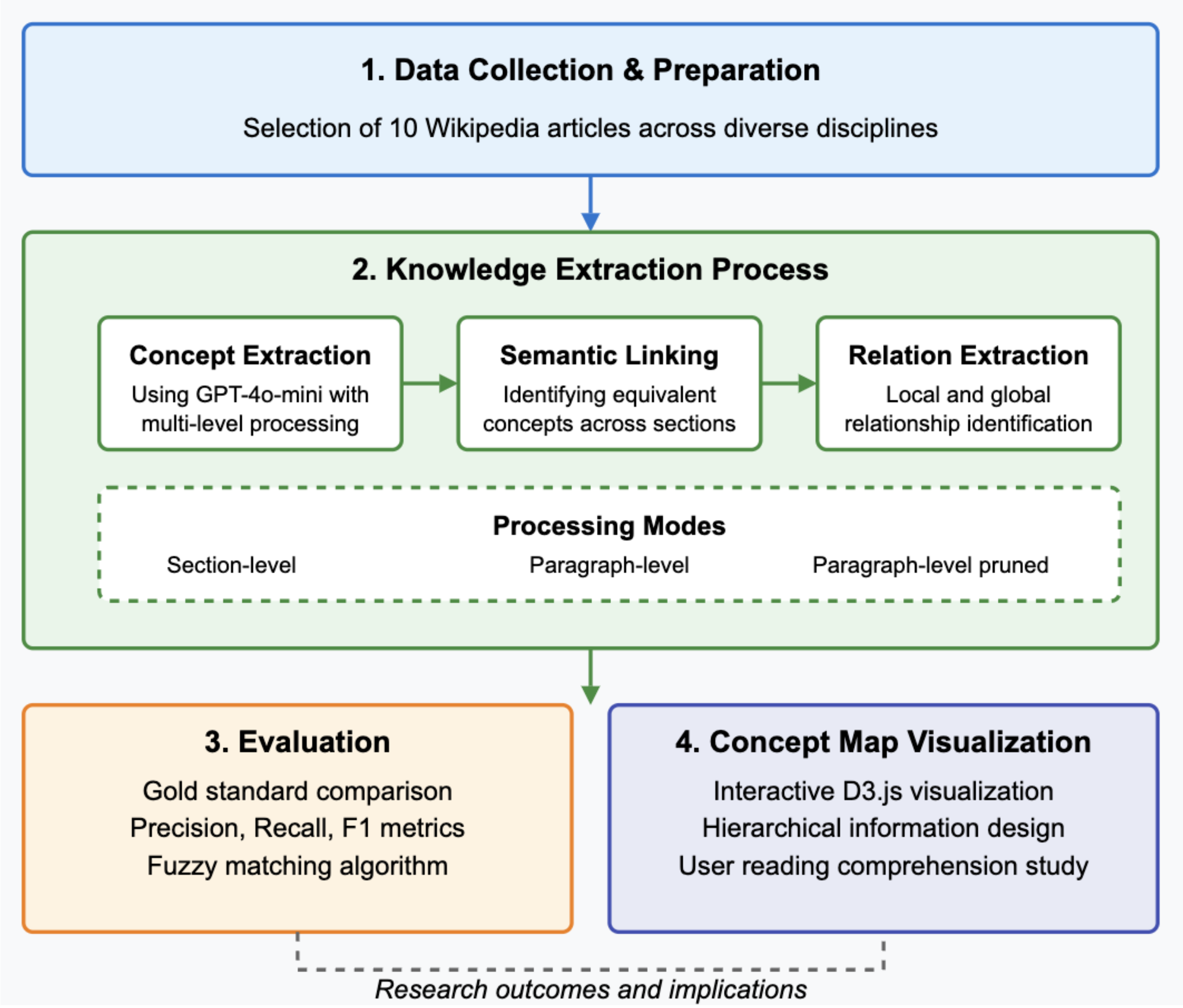

1. Data Collection & Preparation
Selection of 10 Wikipedia articles across diverse disciplines
2. Knowledge Extraction Process
Concept Extraction
Using GPT-4o-mini with multi-level processing
Semantic Linking
Identifying equivalent concepts across sections
Relation Extraction
Local and global relationship identification
Processing Modes
Section-level
Paragraph-level
Paragraph-level pruned
3. Evaluation
Gold standard comparison
Precision, Recall, F1 metrics
Fuzzy matching algorithm
4. Concept Map Visualization
Interactive D3.js visualization
Hierarchical information design
User reading comprehension study
Research outcomes and implications

Figure 3.1: Overview of methodology

blocks, then semantically links the same concepts and extracts relations. It then continues to process subsequent text blocks, performing global relationship extraction after every three text blocks to capture relationships across chapters.

Third, a comprehensive evaluation framework compares the automated extraction results to the gold standard dataset. The comparison utilizes fuzzy string matching to account for small variations in conceptual and relational representations and incorporates standard evaluation metrics including precision, recall, and F1 scores. The evaluation framework allows the system to compare different text processing modalities and assess the extraction performance of the model across disciplines.

Lastly, the concepts and relations extracted by the automated pipeline are transformed into an interactive concept map visualization. This visualization serves as the foundation for evaluating the educational effectiveness of LLM-created concept maps through controlled experimentation. The experimental design assesses reading comprehension outcomes by having participants read two articles matched on length and readability metricsone article without any supplementary aids and the other utilizing solely the generated concept map interface. This controlled comparison enables direct measurement of how concept map-based reading affects comprehension performance, cognitive load, and knowledge retention across diverse academic content.

3.2 Dataset selection and preparation

This project uses Wikipedia articles as the main source of academic content for concept and relation extraction, selecting ten articles across diverse academic disciplines. The selection included articles from biology, mathematics/statistics, computer science, linguistics, art, history, philosophy, political science, health/medicine, and one general non-academic field. Each article was 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 or everyday life. This deliberate focus on unfamiliar content mimics real-life learning scenarios in higher education, where students are regularly required to engage with new and challenging academic material, allowing the testing of extraction methods in contexts that reflect real educational environments.

Each selected article maintained sufficient conceptual depth to support meaningful extraction, a neutral academic tone, and introduced new concepts rather than common basics within its subject area. Wikipedia articles were selected for their free accessibility, consistent structured organization, and extensive study in natural language processing research, while providing comprehensive overviews with depth appropriate for undergraduate understanding.

3.2.1 Preprocessing steps

Before applying the extraction methods, several preprocessing steps were performed on the Wikipedia articles:

1. HTML Removal: All HTML markup and Wikipedia-specific formatting were removed to create clean text documents.

2. Section Identification: The hierarchical structure of each article was preserved by extracting and labeling section headings and subheadings.

3. Reference Removal: In-text citation markers (e.g., [1], [2]) were removed to avoid their interference with the extraction process while preserving the surrounding contextual information.

4. Image and Table Handling: Descriptions of images and tables were retained as text, but the visual elements themselves were excluded from the analysis.

5. Tokenization: The text was tokenized using spaCy's natural language processing library.

No content simplification, summarization, or semantic alteration was performed during preprocessing to maintain the original complexity and depth of the academic content.

3.3 Text processing modes

The automated extraction pipeline implements three distinct text processing modes to evaluate how different granularities of input text affect the quality of concept and relation extraction. These modes provide a systematic framework for comparing extraction performance across varying levels of textual context.

Section-level processing

Section-level processing utilizes complete sections from Wikipedia articles as input units. These sections, naturally delineated by Wikipedia's article structure, typically encompass multiple paragraphs discussing related aspects of a topic. This mode enables the extraction system to process larger chunks of coherent text, potentially capturing broader thematic relationships and high-level concepts that span multiple paragraphs.

Paragraph-level processing

Paragraph-level processing operates on individual paragraphs as discrete input units. This finer granularity allows the extraction system to focus on local concepts and relationships within more concentrated contexts. By processing text at the paragraph level, the system can identify detailed relationships that might be obscured when processing larger sections, while maintaining the natural coherence of ideas typically contained within a single paragraph.

Paragraph-pruned processing

The paragraph-pruned processing mode introduces sophisticated filtering mechanisms to address the limitations of standard paragraph-level processing. This mode applies two complementary filtering approaches to refine extraction outcomes.

For concept pruning, the system eliminates concepts that appear exclusively in single paragraphs, retaining only those with sufficient cross-paragraph relevance. Additionally, the system employs semantic embedding techniques using the all-MiniLM-L6-v2 transformer model to calculate similarity scores between concepts and their section contents, preserving only concepts with similarity scores exceeding 0.6. It was determined through systematic evaluation on a development subset of our corpus, testing values from 0.5 to 0.7 in 0.05 increments. The 0.6 threshold provided the optimal balance between precision and recall across all academic disciplines in our sample, retaining 83% of concepts rated as "primary" or "secondary" while filtering out 76% of concepts rated as "tertiary" in the LLM-extracted results.

For relation pruning, the system implements a dual validation framework. The first validation layer examines the presence of explicit textual evidence supporting each proposed relationship. The second layer calculates a text-to-relation alignment score, which quantifies the degree to which the source text directly supports the extracted relationship. This score must exceed 0.75 to retain the relationship in the final extraction set. This approach distinguishes between substantively meaningful relationships (e.g., “Dogs are mammals,” which directly supports a taxonomic relationship) and mere conceptual co-occurrence without semantic connection. Through these refinement measures, the paragraph-pruned mode achieves extraction quality more comparable to section-level processing while preserving the granular analysis capabilities inherent to paragraph-level processing.

3.4 Model selection

This study employed a combination of models tailored for different aspects of the concept and relation extraction process. The primary extraction model used was GPT-4o-mini, while all-MiniLM-L6-v2 sentence transformer was utilized for paragraph-level pruned filtering.

3.4.1 GPT-4o-mini

GPT-4o-mini was selected as the primary model for concept and relation extraction based on its strong linguistic understanding and ability to identify semantic relationships in text. This model was configured to process academic content and extract both key concepts and the relationships between them. The model’s ability to understand context and identify complex semantic structures made it particularly suitable for processing academic Wikipedia articles. Temperature of responses is set to 0.1 to remain robust results.

```
def _cached_api_call(self, prompt: str) -> str:
    """Cache API calls in memory."""
```

```python
    response = self.client.chat.completions.create(
        model="gpt-4o-mini",
        messages=[{"role": "user", "content": prompt},
                  {"role": "system", "content": "You are an expert at analyzing text and extracting meaningful concepts and relationships between them, with a special focus on making complex information more understandable. "}],
        temperature=0.1
    )
    return response.choices[0].message.content
```

3.4.2 all-MiniLM-L6-v2 sentence transformer

The all-MiniLM-L6-v2 sentence transformer was implemented specifically for paragraph-level pruned filtering, providing an optimal balance between computational efficiency and accuracy for semantic similarity tasks. This model is a distilled version of larger transformer models that has been optimized for creating dense vector representations (embeddings) of text sequences that capture semantic meaning.

The model employs a siamese network architecture with a mean pooling layer, which enables it to generate fixed-size embeddings (384 dimensions) for text sequences of varying lengths. These embeddings position semantically similar sentences closer together in the vector space while placing dissimilar sentences farther apart. This property makes the model particularly well-suited for our similarity comparison tasks between concepts and their section contents.

The model was deployed within our pipeline using the sentence-transformers library, with a cosine similarity threshold of 0.6 established through empirical testing to optimize the balance between precision and recall. This threshold was determined by analyzing the distribution of similarity scores across multiple academic disciplines, identifying the point that maximized the retention of pedagogically relevant concepts while effectively filtering out peripheral or tangential mentions.

3.5 Evaluation framework

To systematically assess the performance of automated concept and relation extraction, this study implemented a comprehensive evaluation framework. This section details the metrics, matching criteria, and analytical approaches used to evaluate extraction quality across different processing modes and academic disciplines.

3.5.1 Quantitative assessment metrics

The evaluation employed three complementary metricsprecision, recall, and F1 scoreeach providing distinct insights into extraction performance:

- **Precision** quantifies the accuracy of the extracted elements by measuring the proportion of correctly identified items among all extractions. It answers the question: "Of all the concepts or relations extracted by the system, what percentage are actually relevant?"

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.1}$$

- **Recall** evaluates comprehensiveness by measuring the proportion of relevant items successfully extracted from the total set present in the gold standard. It answers the question: "Of all the relevant concepts or relations in the gold standard, what percentage did the system successfully extract?"

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.2}$$

- **F1 Score** provides a balanced assessment by calculating the harmonic mean of precision and recall, offering a unified performance metric that considers both dimensions simultaneously. The F1 score is particularly valuable when there is an inherent trade-off between precision and recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

These metrics were calculated separately for both concept and relation extraction.

3.5.2 Matching criteria implementation

To accommodate the linguistic variability inherent in academic text, the evaluation implemented a fuzzy matching algorithm rather than relying solely on exact string matches. This approach recognizes that semantically equivalent concepts or relations may be expressed using different terminology, phrasing, or word order.

The fuzzy matching algorithm employs a hierarchical matching process that begins with strict comparison criteria and progressively applies more flexible matching techniques:

1. **Normalization:** All text strings are standardized by converting to lowercase, trimming extraneous whitespace, and removing common stop words that do not contribute to semantic meaning.

2. **Exact Matching:** The algorithm first attempts to identify exact matches between normalized strings, assigning a perfect similarity score of 1.0 to unambiguous cases.

3. **Edit Distance Matching:** When exact matching fails, the system applies a partial matching technique based on Levenshtein distancea metric that quantifies the minimum number of single-character edits required to transform one string into another. Items with a Levenshtein distance of 2 or less are considered potential matches, accommodating minor spelling variations, pluralization differences, and typographical errors.

4. **Word-level Similarity:** For items that fail both exact and edit-distance matching, the algorithm implements word-level similarity assessment. This approach decomposes multi-word strings into constituent word sets and identifies the proportion of matching words. The similarity score is calculated as the ratio of matching words to the maximum size of either word set.

5. **Score Enhancement:** The algorithm further refines matching precision by applying a score boost for strings demonstrating substantive lexical overlap. When multiple words match between pairs and the base similarity score equals or exceeds 0.5, the algorithm applies a score boost of 0.1.

For relation triplets specifically, this matching process is applied to all three components (source concept, relation type, and target concept), with a relation considered a match only when all components meet the similarity threshold.

The mathematical definition of the similarity function is formalized as follows:

Let s_1, s_2 be two strings after normalization (lowercase, trimmed, stop words removed).

$$\text{Define the similarity function } S(s_1, s_2) \text{ as follows:} \tag{3.4}$$

$$S(s_1, s_2) = \begin{cases} 1.0 & \text{if } s_1 = s_2 \text{ (Exact match)} \\ 0.9 & \text{if } L(s_1, s_2) \leq 2 \text{ (Full string partial match)} \\ \text{word\_similarity}(s_1, s_2) & \text{otherwise (Word-level similarity)} \end{cases} \tag{3.5}$$

$$\text{where word\_similarity}(s_1, s_2) \text{ is defined as:} \tag{3.6}$$

$$\text{Let } W_1, W_2 \text{ be the sets of words in } s_1, s_2 \text{ respectively.} \tag{3.7}$$

$$\text{Let } M = |\{w_1 \in W_1 : \exists w_2 \in W_2, L(w_1, w_2) \leq 2\}| \tag{3.8}$$

$$\text{base\_score} = \frac{M}{\max(|W_1|, |W_2|)} \tag{3.9}$$

$$\text{word\_similarity}(s_1, s_2) = \begin{cases} \min(\text{base\_score} + 0.1, 1.0) & \text{if } M > 1 \text{ and base\_score} \geq 0.5 \\ \text{base\_score} & \text{otherwise} \end{cases} \tag{3.10}$$

Where:

- $L(x, y)$ is the Levenshtein distance between strings
- $|W|$ denotes the cardinality of set W
- M is the number of matching words (including partial matches)

This comprehensive matching approach ensures that the evaluation accounts for the natural variability in how concepts and relations may be expressed, providing a more realistic assessment of extraction performance than would be possible with exact matching alone.

3.6 Knowledge extraction process

3.6.1 Gold standard dataset construction

The creation of reliable gold standard datasets for both concepts and relations followed a systematic manual annotation process. This process established ground truth data against which automated extraction methods could be evaluated. The methodology encompassed initial extraction, structured rating guidelines, independent annotation, inter-annotator agreement assessment, and final dataset reconciliation.

Annotation procedure

The annotation procedure began with a thorough manual extraction of elements from ten academic articles representing different disciplines. Two annotatorsthe primary researcher and a second evaluatorindependently rated these elements according to established guidelines. Both annotators documented their rationales for assigned ratings, particularly for challenging cases requiring careful consideration.

A scale was used to evaluate educational relevance for each element type (concepts and relations). Items receiving a “0” rating were deemed irrelevant or inappropriate for educational concept maps and were excluded from the final gold standard datasets.

Inter-annotator reliability was assessed using Cohen’s Kappa coefficient, which evaluates agreement while accounting for chance. The coefficient is calculated using the formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.11}$$

where P_o represents the observed agreement between annotators, and P_e represents the expected agreement by chance based on the counts from the confusion matrix. The calculation used exact matching criteria, requiring annotators to assign identical ratings for an item to be considered in agreement.

The final datasets were constructed through careful reconciliation of annotator ratings. Annotators reviewed all rating disagreements, with particular attention to cases where one annotator assigned a “0” rating while the other indicated relevance. Following discussion and resolution, all items receiving a consensus “0” rating were systematically removed from the datasets.

Concept annotation guidelines

For the concept annotation process, a concept was defined as “a significant term or phrase that represents a fundamental idea, entity, or phenomenon within a discipline, often reflecting

Rating	Category & Description	Examples
3	**Core Concept**: Essential for basic understanding. Without this concept, reader would fail to grasp the topic.	- Primary definition of article's subject - Fundamental principles - Critical distinctions - Key historical developments
2	**Supporting Concept**: Helpful for deeper comprehension but not essential. Provides context or elaboration.	- Examples illustrating main points - Additional details - Secondary characteristics - Supporting evidence - Contextual information
1	**Peripheral Concept**: Minimally relevant. Could be removed with minimal impact.	- Minor historical details - Tangential facts - Non-critical examples - Slightly related background
0	**Irrelevant Concept**: No direct relationship to main topic. Can be removed without affecting comprehension.	- Unrelated tangents - Excessive technical details - Purely decorative text - Non-contributing information

Table 3.1: Concept rating guidelines

its role in a theory or its relation to other concepts" [16].

The concept rating framework established a 4-point scale for evaluating the pedagogical significance of extracted concepts, as detailed in Table 3.1. This scale provided clear differentiation between concepts based on their importance to understanding the subject matter:

Relation annotation guidelines

For the relation annotation process, a relation was defined as "a meaningful connection between two concepts that expresses a specific type of relationship, such as causation, definition, or composition." The extraction methodology produced triplets consisting of a source concept, a target concept, and the semantic relationship connecting them.

The relation rating framework utilized a similar 3-point scale (0-2) to assess the educational utility of relation triplets, as detailed in Table 3.2:

Rating	Description	Example
2	**Highly Useful and Effective**: Relation is highly relevant to central message or main idea. Simplifies complex ideas into clear connections, providing complete context without requiring additional inferences.	"determiner - adjective - noun pattern" *is an example of* "sentence structure" (Specific, contextual, and directly applicable)
1	**Somewhat Helpful but Needs Improvement**: Relation is somewhat relevant but lacks clarity or context. Provides limited insight without directly assisting understanding of core ideas. May require multiple inferences from the reader.	"sentence parsing" *can use* "serial parsing" (Too vague: doesn't specify how to use or why)
0	**Completely Unhelpful or Harmful**: Relation is irrelevant, confusing, or poorly framed. Lacks clear connection to main ideas, misrepresents information, or introduces unnecessary complexity.	"Digital reading" *has opposite of* "comprehension" (Misrepresents the relationship: digital reading may challenge but doesn't oppose comprehension)

Table 3.2: Relation rating guidelines

This structured annotation process ensured consistent evaluation across different academic disciplines while maintaining focus on pedagogical utility, providing reliable gold standard datasets for both concepts and relations.

3.6.2 Concept extraction

Extraction framework

The concept extraction framework utilizes a pre-trained large language model with precise prompting strategies to identify educationally relevant concepts within academic texts. The extraction process operates on a clearly defined concept framework, where a concept is defined as “a significant term or phrase that represents a fundamental idea, entity, or phenomenon within a discipline.” This definition aligns with established educational theory while providing sufficient specificity for computational extraction. The extraction algorithm processes text at varying granularities (section and paragraph levels) to capture both localized concepts and those that span broader discourse contexts.

Hierarchical extraction parameters

A distinguishing feature of the extraction methodology is its implementation of a hierarchical classification schema that categorizes concepts into three distinct layers based on their educational significance:

- Priority Layer (Core Concepts): This layer encompasses fundamental principles, key terminology, major themes, and critical processes central to understanding the domain. These concepts form the foundation of knowledge representation and constitute approximately 15-20% of extracted concepts.

- Secondary Layer (Supporting Concepts): This intermediate layer includes sub-processes, related theories, component parts, and methodological approaches that elaborate on core concepts. These elements provide essential context and represent 40-50% of the extracted concept base.

- Tertiary Layer (Contextual Elements): This peripheral layer captures author contributions, specific examples, historical developments, applications, and quantitative data that enrich understanding but are not fundamental to core comprehension. These elements constitute 30-40% of the concept base.

This layered approach directly supports the hierarchical information architecture of the visualization system, enabling progressive disclosure of information aligned with cognitive load theories.

Semantic linking process

A critical component of the extraction pipeline is the concept normalization process that identifies semantically equivalent concepts expressed in different linguistic forms. This process employs a specialized comparison algorithm that evaluates potential concept matches based on precise matching criteria:

- Exact Semantic Equivalence: Concepts that refer to identical knowledge units expressed through different terminology.
- Synonymous Expressions: Alternative phrasings that communicate the same underlying concept.
- Contextual Equivalence: Terms that convey the same meaning within the specific domain context.

The algorithm explicitly distinguishes between genuine equivalence and mere relatedness, rejecting matches between:

- Related but distinct concepts (e.g.,“tardigrade anatomy” and “tardigrade”)
- Hierarchically related concepts
- Different aspects of the same broader topic

Key prompt structures

The extraction process implements specific guidelines to ensure balanced coverage across conceptual dimensions. These guidelines explicitly target concepts that answer fundamental knowledge questions:

- “What” concepts (definitions and principles)
- “How” concepts (processes and methods)
- “Why” concepts (reasoning and implications)
- “When” concepts (temporal and contextual factors)

This multidimensional approach ensures that the extracted concept base captures the full range of knowledge types represented in educational texts, from declarative knowledge to procedural and contextual understanding. Below are the key portions of these prompts, with the complete implementations provided in Appendix A.2.

Concept extraction prompt structure The concept extraction prompt defined the fundamental parameters for identifying educationally relevant concepts and provided explicit categorization guidelines:

```
# Key portion of the concept extraction prompt
prompt = f""" 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
   [...]

**Output Format:**
[
    {{
    "entity": "main_form",
    "context": "The exact sentence where this concept appeared",
    "evidence: "Why this concept is essential for understanding the topic",
    "layer": "priority/secondary/tertiary"
    }}
]

Section text:
{full_section_text}
"""
```

Concept linking prompt structure The semantic linking process used a specialized prompt to identify concept equivalence across different sections. This semantic normalization process is implemented through a case-insensitive comparison algorithm with robust exception handling and memory-efficient caching mechanisms to optimize performance with large concept sets. Below is the key prompt structure used in the pipeline:

```
# Key portion of the semantic linking prompt
prompt = f"""
Compare these two lists of concepts and identify which ones represent EXACTLY the same abstract idea or unit of knowledge.

```

```
Guidelines for matching:
1. Match concepts that:
    - Refer to exactly the same concept
    - Are synonyms or alternative expressions
    [...]

2. Do NOT match concepts that:
    - Are merely related or connected (e.g., "tardigrade anatomy" is not equal to "tardigrade")
    [...]

List 1: {json.dumps(normalized_list1, indent=2)}
List 2: {json.dumps(normalized_list2, indent=2)}
"""
```

Technical implementation

The technical implementation incorporates several optimization strategies to ensure efficiency and reliability:

- Multi-level caching: The system implements both memory-based and file-based caching mechanisms to avoid redundant processing of identical text segments or concept comparisons.

- Normalized comparison: All comparisons are performed on lowercase-normalized representations to eliminate false negatives due to case variations while preserving the original case in the final output.

- Structured output format: Extraction results are formatted as structured data objects containing:

- The canonical form of the concept
- The original context (sentence) where the concept appeared
- Supporting evidence explaining the concept for undergraduate-level understanding
- The assigned importance layer (priority, secondary, or tertiary)

3.6.3 Post concept extraction analysis

Concept categorization framework

The categorization of extracted concepts follows a structured framework adapted from recent work in Concept and Named Entity Recognition (CNER) by Martinelli et al.[36]. While traditional NER and concept extraction focuses on identifying general entities and concepts, our methodology is specifically tailored for educational concept mapping. This educational focus influences both the extraction and categorization processes, as concepts are selected and organized with the goal of supporting visual, non-linear learning experiences.

The categorization framework comprises seventeen distinct categories, some aligning with traditional CNER categories (e.g., PERSON, ORGANIZATION, LOCATION) while others are specifically designed for educational concept mapping. Core Concepts form the foundation of subject matter understanding, while categories such as Research & Analysis, Processes, and Classification capture the methodological and organizational aspects of academic knowledge. Categories like Socio-Cultural Contexts and Historical Events reflect the broader contextual elements essential for comprehensive understanding.

Table 3.3 presents the complete categorization framework, with each category including a description and representative examples from the extracted concepts. The framework's structure reflects the multifaceted nature of academic knowledge, encompassing not only fundamental disciplinary concepts but also their applications, contexts, and real-world implications. This comprehensive categorization supports the creation of layered visualizations that enable students to explore connections between different types of concepts.

CNER	**Category**	**Description**	**Examples**
DISCIPLINE	Core Concepts	Fundamental terms and ideas that define the subject matter	*roanoke colony, consociationalism, chaos theory*
-	Research & Analysis	Methods and studies used to analyze or understand the subject	*empirical average-case complexity, genome size variation, psycholinguistic research*
CULTURE	Socio-Cultural Contexts	Social and cultural influences or contexts that shape the subject	*japanese culture, cultural symbolism, language shaping*
-	Processes, Mechanisms, and Procedural Elements	Processes, methods, or mechanisms involved in the subject	*symbiosis, post-conflict state-building, bifurcation*
-	Classification and Taxonomy	Systems for categorizing or organizing concepts within the subject	*polynomial time hierarchy, class eutardigrada, taxonomy of tardigrades*
EVENT & DATETIME	Historical Events and Evolution	Important historical moments, shifts, or movements within the subject	*samuel mace's 1602 voyage, post-war disenchantment, fossil record*
STRUCT & PART	Structural Features	Specific structural components or parts that make up the subject	*nervous system, cold chambers, structural causes of epistemic injustice*
PROPERTY	Properties and Attributes	Characteristics or qualities inherent in the subject matter	*credibility, pain intensity, simplicity*
LOC	Environmental Contexts	Environmental or contextual factors that influence the subject	*habitat specialization, st. john's, Newfoundland, port ferdinando*
PER & ORG	Key People and Organizations	Important figures and organizations that contribute to the subject	*kurt gdel, international actors, gutai group*
ARTIFACT	Documents and Artifacts	Physical or digital artifacts that are relevant to the subject	*fluxkits, readymades, scott dawson's book, gilbert and lila silverman collection*
-	Problems and Solutions	Challenges and resolutions related to the subject matter	*generalized sudoku problem, two-state solution, act-out errors*
-	Mathematical and Computational Foundations	Theoretical or computational bases that support the subject	*polynomial time, rotation number, big o notation*
-	Applications and Real-World Impact	Practical applications or real-world effects of the subject	*cryptography impact, legacy of the colony, epistemic injustice in health*
LAW & MONEY & ASSET	Political and Economic Systems	Political or economic structures and stances related to the subject	*economic motivations, proportional employment, political stance*
DISEASE & SUBSTANCE	Medical and Safety Considerations	Health, safety, and medical implications related to the subject	*neglected tropical disease, safety precautions, hypnosis*
MEDIA	Media	Forms of media used to express or document the subject	*video art, media adaptations, the Simpsons - treehouse of horror vi*

Table 3.3: Label, description, and examples of each concept category.

Quantitative assessment against gold standard

The primary quantitative evaluation employed is described in three complementary metricsprecision, recall, and F1 score. To accommodate the linguistic variability inherent in academic text, the evaluation implemented fuzzy matching criteria, detailed in Section 3.5. The fuzzy matching algorithm employs a hierarchical matching process that begins with strict comparison criteria and progressively applies more flexible matching techniques.

3.6.4 Relation extraction

Relation extraction framework

The relation extraction framework utilizes natural language processing techniques through large language models to identify semantically meaningful connections between previously extracted concepts. The system defines relations as structured triplets consisting of a source concept, a target concept, and a descriptive relation type that characterizes the semantic connection between them. Each relation is further contextualized with supporting evidence from the source text to validate its authenticity and enhance educational comprehension. The extraction architecture employs a multi-tiered approach that distinguishes between two complementary types of relationships:

- Local Relations: Connections between concepts that occur within the same textual segment (section or paragraph), representing immediate semantic relationships established within a specific context.

- Global Relations: Higher-order connections between concepts that may not co-occur directly but demonstrate significant relationships across different sections of the document, representing broader thematic or logical associations.

This dual-layer approach ensures comprehensive coverage of the conceptual relationships, from granular, context-specific associations to overarching structural patterns that define the knowledge domain.

Key prompt structures

The extraction process employed carefully crafted prompts to instruct the language model in identifying relations. Below are the key portions of these prompts, with the complete implementations provided in Appendix A.2.

Local relation extraction methodology The local relation extraction process operates at the section level, examining semantic connections between concepts that co-occur within the same textual unit. This approach recognizes that relationships often manifest within cohesive discourse segments where related concepts are naturally introduced and explained in proximity to one another. The extraction algorithm employs precisely formulated prompts that instruct the language model to:

- Identify clear and well-defined relationships between available concepts within the section
- Capture diverse relationship types without imposing restrictive categorical limitations
- Focus exclusively on relationships with explicit textual support or strong inferential evidence
- Provide specific textual evidence for each proposed relationship to validate its inclusion

```
# Key portion of the local relation extraction prompt
prompt = f"""
Extract key relationships between these available concepts using the following guidelines.

**Context:**
The extracted relations should represent meaningful connections that contribute to understanding the main ideas in the text.

```

```
**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.
[...]

Available Concepts: {json.dumps([c["id"] for c in concepts], indent=2)}
Section Text: {text}
"""
```

Global relation extraction methodology Complementing the local extraction process, the global relation methodology identifies higher-order connections that span across different sections of the text. This approach recognizes that significant conceptual relationships often transcend immediate textual proximity, especially in complex educational materials where key concepts may be revisited and interconnected throughout the document. The global extraction algorithm utilizes specialized prompts that direct the language model to:

- Identify relationships of significance at a higher structural level beyond individual sections
- Detect patterns of conceptual influence across different contexts within the document
- Focus on overarching connections that enhance holistic comprehension of the material
- Provide reasoned justification for each proposed global relationship to establish validity

```
# Key portion of the global relation extraction prompt
prompt = f"""
```

```
Extract global relationships using all processed concepts. The focus is on identifying high-level connections that span across sections or paragraphs.

**Context:**
The extracted global relationships should illustrate overarching connections that tie together multiple sections.

**Guidelines:**
- Identify relationships that are significant at a higher level, beyond individual sections or paragraphs.
- Include relationships that show how concepts influence each other across different contexts or sections.
[...]

Available Concepts: {json.dumps([c["id"] for c in master_concepts], indent=2)}
"""
```

Evidence-based validation

A critical feature of both local and global relation extraction processes is the systematic collection of supporting evidence. For each identified relationship, the system extracts or generates a concise explanation that:

- References specific textual content supporting the relationship
- Articulates the nature of the semantic connection between the concepts
- Contextualizes the relationship within the broader knowledge framework

This evidence-based approach serves multiple purposes: it validates the authenticity of extracted relationships, provides educational context that enhances learner comprehension, and offers transparency regarding the extraction rationale.

Technical implementation

The technical implementation incorporates several optimization strategies to ensure efficiency and reliability:

- Multi-Level Caching: Both memory-based and file-based caching mechanisms are implemented to prevent redundant processing of identical text segments or concept combinations. This optimization significantly reduces computational overhead during iterative processing cycles.

- Structured Data Management: All extracted relationships are systematically stored in a structured format that preserves the source and target concepts, relation type, supporting evidence, and section context. This organization facilitates subsequent retrieval, analysis, and visualization.

- Error Handling: Robust exception management mechanisms ensure that extraction failures for individual relationships do not compromise the overall extraction process, maintaining system stability throughout complex document processing.

- Comprehensive Relation Tracking: The system maintains three distinct relation collections: local relations (section-specific), global relations (document-spanning), and master relations (the complete consolidated set), enabling flexible access to relationship data at different granularity levels.

3.6.5 Post relation extraction analysis

Relation categorization framework

To systematically analyze the semantic connections extracted from educational texts, we developed a comprehensive relation categorization framework (Table 3.4). This framework, adapted from Asher's work on discourse relations [3], classifies relationships into seven distinct categories based on their semantic function. Each category captures a fundamental type of connection between concepts, from structural composition to causal mechanisms and functional purposes. It's worth noting that the "Causal (expanded)" category incorporates evidence relations, as evidence typically indicates or supports causal relationships between concepts. This classification enables both quantitative analysis of relation distribution across academic disciplines and qualitative assessment of how different domains structure knowledge. By applying this framework consistently across all extracted relations, we can identify discipline-specific patterns in conceptual organization while maintaining comparability across domains.

Quantitative assessment against gold standard

The performance evaluation of relation extraction uses the same three core metrics as described in 3.5. However, these metrics reflect the additional complexity of evaluating structured triplets rather than individual entities.

For evaluating relation extraction, the fuzzy matching algorithm was applied comprehensively to all components of relation tripletssource concept, target concept, and relation type. The evaluation calculates separate metrics for source concept matching, target concept matching, relation type matching, and complete triplet matching. This ensures that partially correct relations (e.g., correct concepts but incorrect relationship) are appropriately differentiated from completely correct extractions.

Relation Type	Description	Examples
Structural	Part-whole relationships, membership, composition, physical/conceptual structure	"contains", "includes", "consists of", "comprises"
Causal(expand)	Direct causation, results/outcomes, supporting evidence, justification	"results in", "provides evidence for", "triggers", "produces"
Impact	Influence patterns, effects, cultural/societal impact	"influences", "affects", "impacts"
Functional	Purpose and usage, capabilities, environmental functions, location-based functions	"used for", "enables", "supports", "applies to"
Interaction	Bidirectional relationships, mutual influences, system interconnections	"relates to", "overlaps with", "interacts with"
Attribution	Origin, creation, theoretical foundations, credit/authorship	"theorized by", "coined by", "authored by", "created by"
Exemplification	Instances, demonstrations, case studies	"exemplifies", "demonstrates", "cases include"
Temporal	Time-based relationships, sequences, chronology, evolution	"precedes", "follows", "occurs during", "evolves into"
Cognitive	Mental processes and understanding, involving mental operations, processing and comprehension, learning and analysis	"involves cognitive mechanisms", "requires analysis", "processes information", "facilitates understanding"
Linguistic	Language-specific patterns and structures, grammatical or syntactic features, language elements and rules, verbal expression	"related to agent-action-patient structure", "has syntactic property", "follows grammatical pattern", "exhibits linguistic feature"

Table 3.4: Summary of relation types, descriptions, and examples

3.7 Full automated extraction workflow

The automated extraction process implements a systematic workflow to identify concepts and relations from academic texts using GPT-4o-mini. Figure 3.2 illustrates this workflow, which processes text sections sequentially while maintaining connections across the entire document.

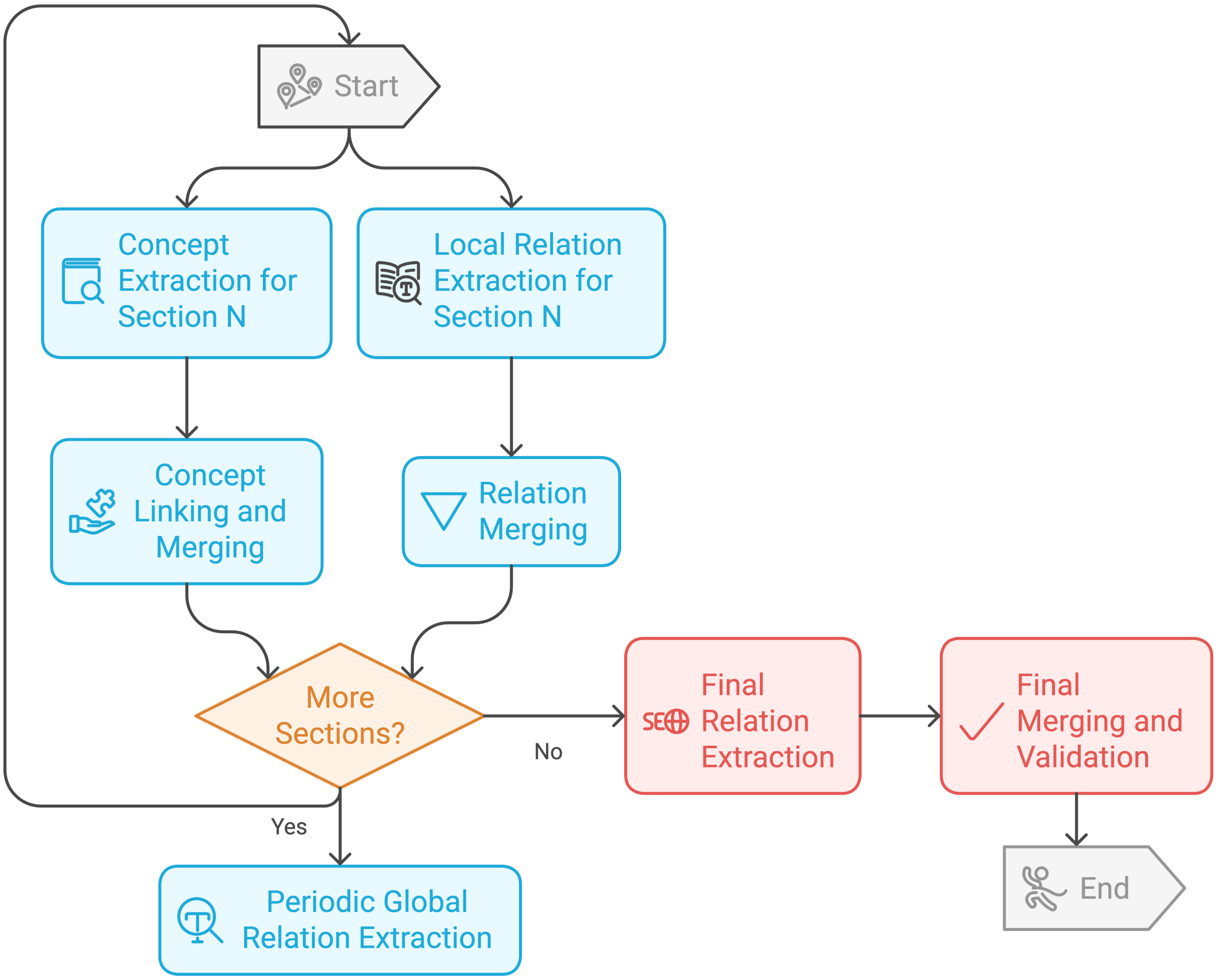

Start
Concept Extraction for Section N
Local Relation Extraction for Section N
Concept Linking and Merging
Relation Merging
More Sections?
No
Yes
Final Relation Extraction
Final Merging and Validation
Periodic Global Relation Extraction
End

Figure 3.2: Concepts and relations extraction workflow

The workflow begins with concept extraction from the first section, immediately followed by relation extraction for the same section. After processing two sections, the system implements concept linking and merging to identify semantically identical concepts across sections and consolidate them, ensuring consistent representation throughout the document. Relation merging follows, combining similar relationships while preserving their unique contextual meanings.

A crucial feature of this workflow is the global relation extraction phase, which occurs after processing multiple sections. This phase examines the entire set of previously extracted concepts to identify relationships that span across sections, capturing higher-level connections not apparent within individual sections. This dual approach to relation extractionboth local within sections and global across sectionsensures comprehensive capture of relationships at all levels of textual organization.

The workflow concludes with a final merging and validation phase, consolidating all extracted information and ensuring consistency in the final concept and relation sets. This systematic approach enables thorough concept and relation extraction while maintaining the coherence of the document's conceptual structure.

3.8 Concept map visualization

3.8.1 Concept map construction

The implementation of concept maps in this study followed established design principles from conceptual visualization research (ConceptEVA) [59] and Novak's foundational work on concept mapping [40]. The concept map was constructed with several essential features designed to optimize the user experience while maintaining educational effectiveness.

The concept map visualization was designed with five integrated features that work together to enhance the educational experience while managing cognitive load. These features collectively support a scaffolded, user-directed learning experience:

Hierarchical information architecture. A layered display approach was implemented to prevent cognitive overload while maintaining access to comprehensive information. The default view presented only concepts from the priority layer, representing the most foundational ideas necessary for understanding the core content. This directly leveraged the layered information generated during the concept extraction phase, where each concept was systematically categorized based on its importance. The architecture enabled progressive disclosure of information, allowing users to construct mental models incrementally while preserving all underlying connections between concept levels.

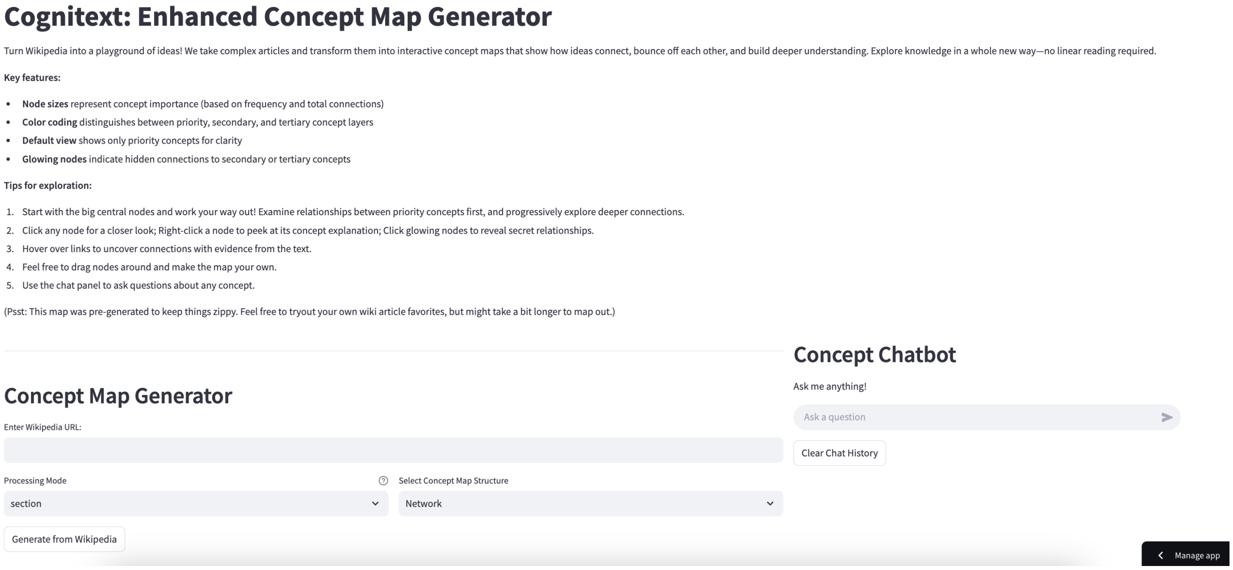
Self-directed exploration. Building upon the hierarchical architecture, the interface promoted active learning through user-driven exploration. Nodes featured visual indicators of hidden connectionscues that additional relationships existed with secondary or tertiary concepts not currently visible. These indicators served as implicit invitations for further exploration, encouraging users to expand nodes of interest to discover additional conceptual relationships. This approach aligned with constructivist learning principles, supporting diverse learning approaches and accommodating individual differences in background knowledge.

Visual focus management. To maintain clarity when exploring complex concept networks, the interface implemented a focus management system where selecting a concept node brought that node and its related concepts to the foreground while fading unrelated concepts into the background. This created a temporary subgraph centered on the concept of interest, reducing distractions while helping users concentrate on relevant information. The approach mitigated the complexity of densely connected concept networks while preserving awareness of the overall knowledge structure.

Relationship transparency. The interface revealed detailed relationship information when users hovered over links between nodes, including the specific relationship type and supporting evidence from the source text. This feature promoted understanding of not just what concepts were related, but how and why they were interconnected. This transparency was particularly important for educational applications, as it made explicit the reasoning behind conceptual connections that might otherwise remain implicit in linear text.

Intelligent content enhancement. The concept map was augmented with large language model capabilities to provide contextual explanations for deeper comprehension. Right-clicking on a concept node displayed an LLM-generated explanation calibrated for undergraduate-level understanding. Additionally, a specialized chatbot function was implemented in the right panel of the interface, designed to answer questions based specifically on the generated concept map rather than drawing from general knowledge. This constraint ensured information consistency with the concept map and original text, avoiding potential contradictions that might confuse learners.

These design features work together as an integrated system. This comprehensive approach addresses the challenges of digital reading comprehension by transforming linear text into an interactive knowledge structure that supports diverse learning needs and preferences. Figure 3.3 shows the interface of the Cognitext webapp and Figure 3.4 demonstrates an example network concept map of the article Quantum Supremacy.



Cognitext: Enhanced Concept Map Generator

Turn Wikipedia into a playground of ideas! We take complex articles and transform them into interactive concept maps that show how ideas connect, bounce off each other, and build deeper understanding. Explore knowledge in a whole new way—no linear reading required.

Key features:

- **Node sizes** represent concept importance (based on frequency and total connections)
- **Color coding** distinguishes between priority, secondary, and tertiary concept layers
- **Default view** shows only priority concepts for clarity
- **Glowing nodes** indicate hidden connections to secondary or tertiary concepts

Tips for exploration:

1. Start with the big central nodes and work your way out! Examine relationships between priority concepts first, and progressively explore deeper connections.
2. Click any node for a closer look; Right-click a node to peek at its concept explanation; Click glowing nodes to reveal secret relationships.
3. Hover over links to uncover connections with evidence from the text.
4. Feel free to drag nodes around and make the map your own.
5. Use the chat panel to ask questions about any concept.

(Psst: This map was pre-generated to keep things zippy. Feel free to tryout your own wiki article favorites, but might take a bit longer to map out.)

Concept Map Generator

Enter Wikipedia URL:

Processing Mode: section

Select Concept Map Structure: Network

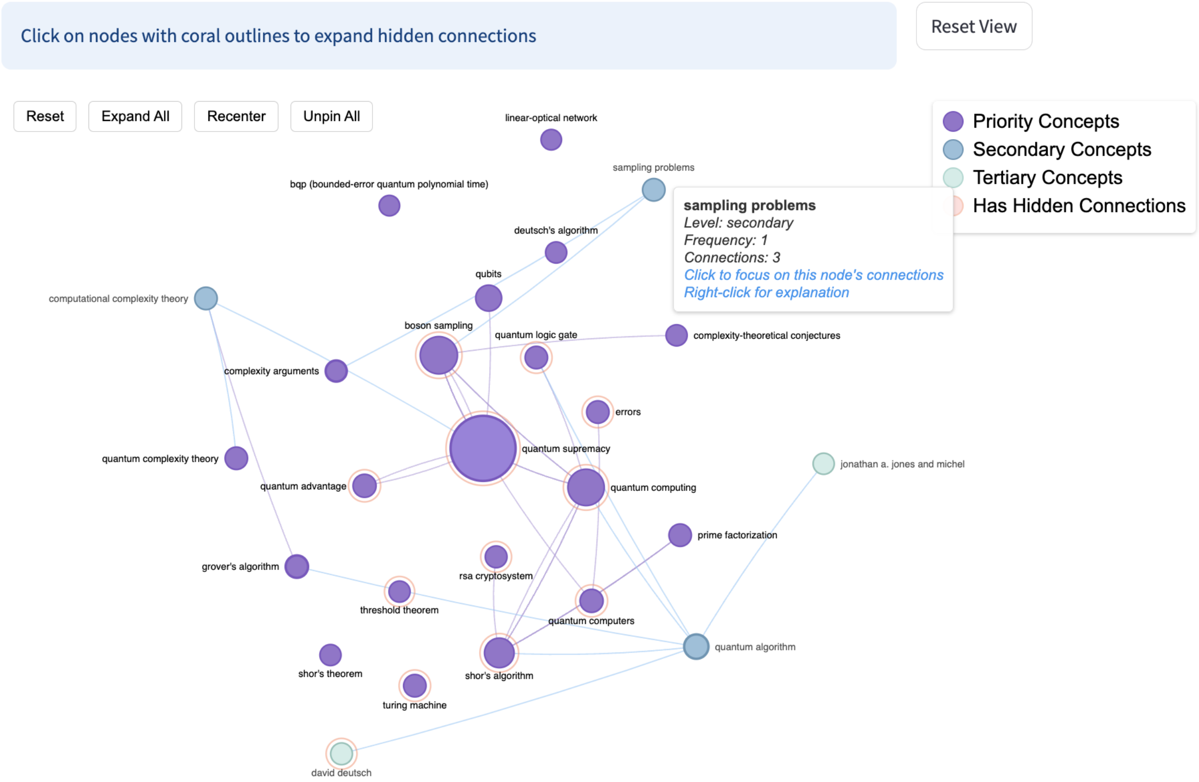
Concept Chatbot

Ask me anything!

Figure 3.3: Cognitext webapp interface

3.8.2 Technical implementation

The concept map visualization was implemented using D3.js for interactive data visualizations, with a Python data processing pipeline handling the extraction and organization of concepts and relationships. The complete interface was deployed through Streamlit Cloud, providing a responsive web application accessible across devices without requiring local installation. This architecture ensured scalability and accessibility while maintaining the interactive performance necessary for fluid concept exploration.



Click on nodes with coral outlines to expand hidden connections

sampling problems
Level: secondary
Frequency: 1
Connections: 3
Click to focus on this node's connections
Right-click for explanation

Figure 3.4: Example network map of the article Quantum Supremacy

3.8.3 User reading comprehension assessment

To evaluate the efficacy of concept maps as visual aids for reading comprehension, a controlled assessment protocol was implemented comparing traditional linear reading with concept map-assisted reading. This assessment targeted multiple dimensions of reading comprehension, including factual recall, conceptual understanding, relational knowledge, and knowledge transfer.

Assessment design

Two academic articles of comparable complexity were selected for the assessment, carefully matched on length and readability as determined by Flesch-Kincaid readability scores and validated through large language model evaluation. This matching process ensured that differences in comprehension outcomes could be attributed to the reading method rather than variability in the reading materials themselves. Participants were required to read one article using the traditional linear approach without assistance, and the other article exclusively through the concept map interface. To mitigate order effects, the sequence of articles and reading methods was randomized across participants. This counterbalanced design controlled for potential learning effects and ensured that results were not influenced by the order of exposure to either reading method.

Comprehension measures

Following each reading task, participants completed a comprehensive assessment consisting of seven questions designed to measure different aspects of comprehension:

- Factual Recall (3 questions): Multiple-choice questions targeting specific factual information presented in the article. These questions were manually constructed to focus on key information points, while response options were generated by a large language model to ensure standardized difficulty levels.

- Conceptual Knowledge (2 questions): Short-answer questions requiring participants to explain core concepts presented in the article, demonstrating deeper understanding beyond surface-level recall.

- Relational Identification (1 question): A structured question requiring participants to identify and explain relationships between key concepts, measuring their grasp of the logical connections within the text.

- Knowledge Transfer (1 question): A novel scenario-based question designed to assess participants' ability to apply newly acquired knowledge to an unfamiliar context, demonstrating higher-order cognitive processing.

Performance metrics

The assessment protocol incorporated multiple metrics to provide a comprehensive evaluation of the reading experience:

- Temporal Efficiency: Participants recorded their total reading time and assessment completion time for each article, allowing for analysis of the relative efficiency of each reading method.

- Cognitive Load: Following each reading and assessment session, participants rated their perceived mental effort using Hart and Staveland's NASA Task Load Index (TLX).

- Comprehension Accuracy: Responses to assessment questions were scored according to a standardized rubric that differentiated between complete, partial, and incorrect answers.

Qualitative feedback

In addition to quantitative measures, participants provided structured feedback regarding their experience with each reading method. This qualitative component included reflections on:

- The overall user experience with both reading approaches

- Specific beneficial aspects of the concept map interface

- Challenges encountered during concept map navigation

- Suggested improvements for the concept mapping tool

This combination of quantitative performance metrics and qualitative user feedback provided a holistic evaluation of concept maps as a reading comprehension aid in educational contexts. The assessment was designed to detect potential differences not only in terms of comprehension outcomes but also in terms of cognitive efficiency and user satisfactioncritical factors for the practical application of such tools in educational settings.

Chapter 4

Results

4.1 Data section

4.1.1 Data characteristics

The Wikipedia articles selected for this study provided suitable test cases for concept and relation extraction across diverse academic domains:

1. **Content diversity**: The corpus spanned ten distinct academic disciplines, ensuring broad representation across STEM fields, humanities, and social sciences. Each article represented specialized knowledge that undergraduate students would typically encounter but might not be familiar with from prior education.

2. **Length variation**: As shown in Table 4.1, the articles ranged from 1,383 to 11,337 words (mean: 4,839), with corresponding token counts from 2,096 to 14,833 (mean: 6,559). This variation allowed us to assess extraction performance across texts of different complexities while remaining comparable to typical undergraduate reading assignments.

3. **Structural consistency**: Each article followed Wikipedia's standardized format with an introduction, hierarchical subsections, and consistent citation practices. This struc-

tural uniformity provided controlled conditions for comparing extraction performance across disciplines while maintaining ecological validity.

4. **Concept density**: Manual annotation revealed an average of 45-60 distinct academic concepts per article, with disciplinary variations reflecting different knowledge organization patterns. This density provided sufficient conceptual material for meaningful extraction while remaining manageable for comprehensive analysis.

5. **Academic complexity**: Flesch-Kincaid readability scores ranged from 30-50, characteristic of undergraduate textbooks. This college-level reading difficulty ensured the corpus represented authentic academic discourse rather than simplified content.

Title	**Category**	**Word Count**	**Token Count**
P versus NP problem	Computer Science	6,042	8,043
Tardigrades	Biology	3,413	5,037
Lost Colony of Roanoke	History	11,337	14,833
Epistemic injustice	Philosophy	1,383	2,096
Consociationalism	Political Science	2,748	3,714
Garden path sentence	Linguistics	2,274	2,782
Fluxus	Arts	7,733	10,524
Mandelbrot set	Mathematics/Statistics	5,850	9,243
Cryotherapy	Health/Medicine	2,328	2,877
Mandela effect	General	5,286	6,441
Average		**4,839**	**6,559**

Table 4.1: Word and token counts by article

4.2 Concept extraction performance by discipline

To establish the reliability of our manually annotated gold standard datasets, we evaluated the level of agreement between two independent annotators who assessed both concepts and relations according to the rating guidelines described in sections 3.6.1 and 4.4. Cohen's Kappa coefficient was calculated to measure inter-annotator agreement while accounting for chance agreement.

For concepts, the confusion matrix in Table 4.2 shows the distribution of ratings between Annotator 1 (rows) and Annotator 2 (columns). Out of 726 total concepts evaluated, the observed agreement proportion (po) was calculated at 0.8526, indicating that annotators agreed on 85.26% of all concept assessments. The expected agreement proportion (pe) was determined to be 0.3773, representing the agreement expected by chance. The resulting Cohen's Kappa coefficient ((κ)) was 0.7633, indicating substantial inter-annotator agreement and attesting to the reliability of our concept assessment guidelines.

Rater 1 \Rater 2	**0**	**1**	**2**	**3**
0	8	0	0	0
1	0	185	44	0
2	0	32	310	31
3	0	0	0	116

Table 4.2: Concept rating agreement

4.2.1 Concept categorization results

The distribution of concept types across different disciplines reveals distinctive patterns that reflect the unique focus and nature of each academic domain. To facilitate meaningful cross-disciplinary comparison of concept distribution patterns, I implemented a normalization procedure that accounts for the substantial variation in text length across the analyzed articles.

The normalization process involved calculating the frequency of concepts per 1,000 words for each article, using the formula:

$$\text{Normalized Count} = \frac{\text{Raw Concept Count}}{\text{Word Count}} \times 1,000 \tag{4.1}$$

Figure 4.1 presents a heatmap visualization of the normalized concept distribution across academic disciplines (shown numerically in Table A.2). The discipline labels (CS, BIO, etc.) represent the individual articles selected from each discipline, rather than generalizable patterns of the entire disciplines. Color intensity represents the frequency of concepts per 1,000 words, with darker blue indicating higher density. This visualization highlights both discipline-specific concept patterns and cross-disciplinary similarities that will be examined in detail in the analysis section.

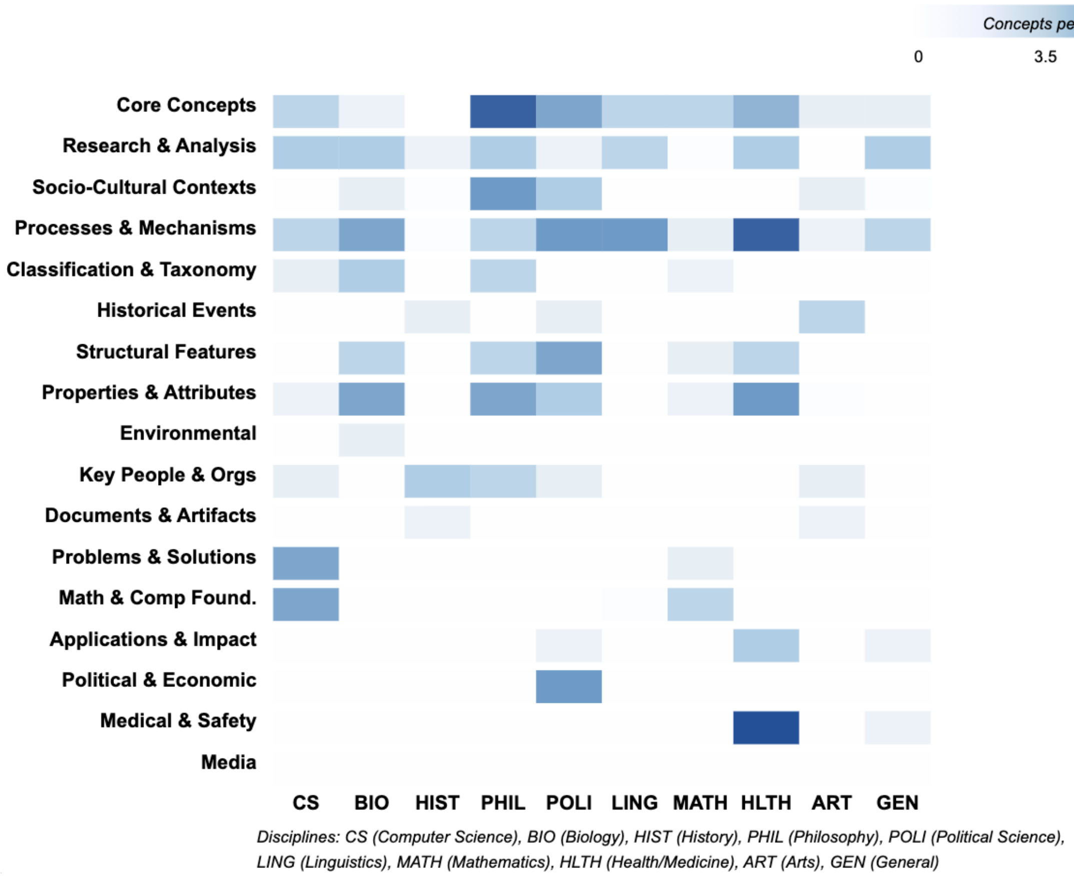


Figure 4.1: Concept distribution heatmap across academic disciplines

When normalized, the Philosophy article showed the highest concept density (31.09 concepts per 1,000 words), followed closely by Health/Medicine (30.93) and Political Science (27.66). In contrast, the History (9.61) and Art (10.86) articles exhibited the lowest concept density despite having high raw concept counts. Computer Science and Biology articles showed similar moderate densities (19.03 and 19.63, respectively).

Normalization also highlighted discipline-specific patterns in concept distribution. In the Computer Science article, Problems & Solutions and Mathematical & Computational

Foundations both achieved the highest normalized frequency (3.14 concepts per 1,000 words). The Health/Medicine article showed a pronounced emphasis on Medical & Safety concepts (7.73) and Processes & Mechanisms (6.87). The Philosophy article demonstrated a substantial focus on Core Concepts (6.51) and Socio-Cultural Contexts (4.34).

The highest normalized values for individual concept categories were observed in the Health/Medicine (Medical & Safety: 7.73), Philosophy (Core Concepts: 6.51), and Linguistics articles (Processes & Mechanisms: 4.84). The History article, despite having the highest raw count for Key People & Organizations (27), showed a more modest normalized value (2.38) due to its greater text length.

Processes & Mechanisms emerged as the most consistently represented category across disciplines after normalization, with notable presence in the Health/Medicine (6.87), Linguistics (4.84), and Political Science (4.37) articles. Some categories remained discipline-specific even after normalization, with Mathematical & Computational Foundations appearing almost exclusively in the Computer Science (3.14) and Mathematics (2.05) articles, and Political & Economic concepts concentrated in the Political Science article(4.00).

4.2.2 Concept extraction performance results

Table 4.3 presents the detailed performance metrics for concept extraction across various academic disciplines using our fuzzy matching algorithm. It's important to note that these results represent performance on single representative articles from each discipline. The evaluation compares three distinct text processing approaches: section-level, paragraph-level, and paragraph-level pruned processing.

The section-level processing approach demonstrated superior precision across all disciplines, achieving an average precision of 83.62%. The Biology article exhibited the highest precision at 89.86%, followed by the General domain text at 87.35% and the History article at 85.47%. However, section-level processing showed comparatively lower recall (62.18% on average), indicating that while this approach extracted highly relevant concepts, it missed a substantial portion of concepts present in the gold standard dataset.

In contrast, paragraph-level processing yielded considerably higher recall metrics across all disciplines, with an average recall of 74.51%. The highest recall values were observed in the Political Science (81.63%), History (81.35%), and Biology (79.62%) articles. This improvement in recall came at the cost of precision, which dropped to an average of 57.49%, substantially lower than the section-level approach. This trade-off suggests that processing text at a finer granularity captures more concepts but introduces more false positives.

The paragraph-level pruned approach attempted to balance these trade-offs, achieving intermediate performance in both precision (66.87% average) and recall (70.92% average). This approach showed the most balanced performance across disciplines, with the General domain exhibiting the highest F1 score of 73.25%, followed closely by the History article at 73.21% and the Biology article at 72.72%.

When comparing F1 scores, which provide a balanced measure of precision and recall, the section-level approach performed best overall with an average F1 score of 71.20%. The paragraph-level pruned approach followed closely with an average F1 score of 68.82%, while the standard paragraph-level approach achieved 64.89%.

Notably, the Linguistics article consistently showed the lowest performance across all processing approaches, with F1 scores of 61.87% (section-level), 56.17% (paragraph-level), and 61.88% (paragraph-level pruned). This indicates that linguistic texts present unique challenges for concept extraction, possibly due to their abstract nature or specialized terminology.

4.3 Relation extraction performance by discipline

For relation triplets, Table 4.4 presents the confusion matrix across all disciplines. Out of 1,139 total relation triplets evaluated, the observed agreement proportion (P_o) was calculated at 0.8165, indicating that annotators agreed on 81.65% of all relation assessments. The expected agreement proportion (P_e) was determined to be 0.3610, resulting in a Cohen's

Discipline	Section-Level			Paragraph-Level			Paragraph-Level Pruned		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	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 4.3: Performance of concept extraction by discipline with fuzzy matching

Kappa coefficient (*kappa*) of 0.7128.

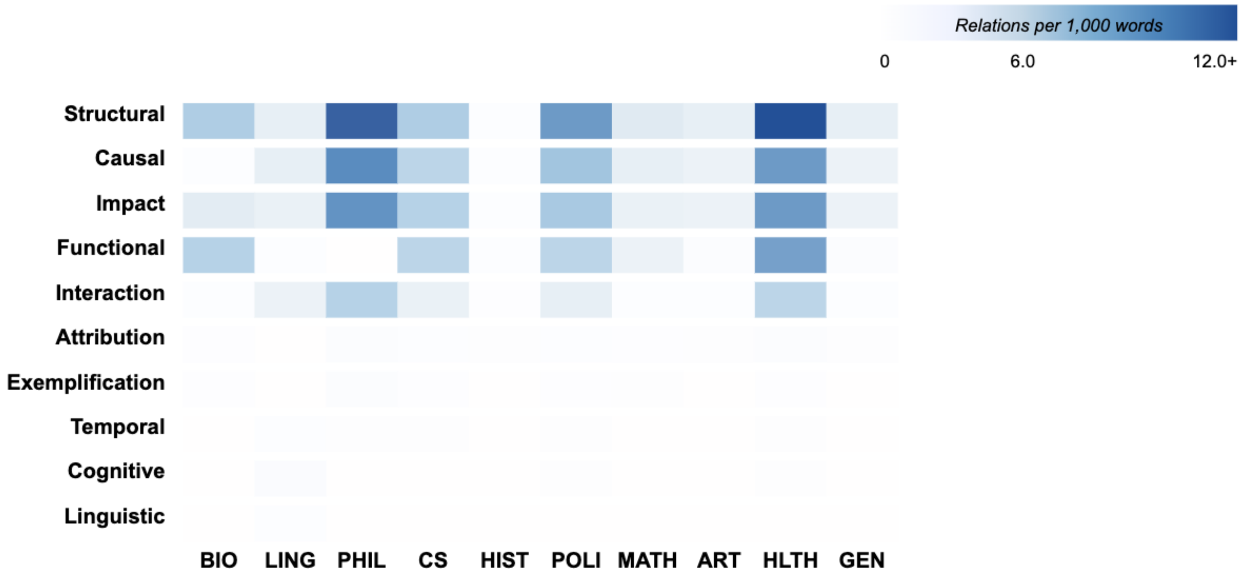
Rater 1 \Rater 2	**0**	**1**	**2**
0	260	43	0
1	48	430	55
2	0	63	240

Table 4.4: Relation rating agreement

4.3.1 Relation categorization results

To visualize the patterns in relation distribution, Figure 4.2 presents a heatmap of the normalized relation frequencies across disciplines (corresponding to the data in Table A.4). The color intensity represents relations per 1,000 words, with the darkest cells indicating the highest densities. These patterns reflect characteristics of the individual articles selected from each discipline rather than definitive disciplinary patterns. This visualization immediately makes apparent both the dominance of structural relations across disciplines and the distinctive relationship signatures of different domains.

The overall relation density varied substantially across disciplines. The Health/Medicine article exhibited the highest density with 54.22 relations per 1,000 words, followed by the



Disciplines: BIO (Biology), LING (Linguistics), PHIL (Philosophy), CS (Computer Science), HIST (History), POLI (Political Science), MATH (Mathematics), ART (Arts), HLTH (Health/Medicine), GEN (General)

Figure 4.2: Relation distribution heatmap across academic disciplines

Philosophy (45.55) and Political Science (41.48) articles. In contrast, the History text showed the lowest relation density (11.47) despite having the third-highest raw count of relations. This inverse relationship between raw counts and normalized density reflects the substantial differences in text length across disciplines.

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. This consistency suggests that taxonomic and compositional relationships form a fundamental aspect of knowledge organization regardless of domain.

Causal and Impact relations showed similar distribution patterns, with the Philosophy text demonstrating remarkably high densities in both categories (10.85 and 10.12 respectively). The Health/Medicine article also exhibited a strong representation of these relation types (10.31 and 9.88).

Specialized relation types revealed distinctive disciplinary signatures. The Linguistics text showed the highest normalized values for domain-specific relations including Cognitive (3.08), Temporal (2.20), and Linguistic (2.20) categories. The Philosophy article demonstrated a complete absence of Functional relations (0.00) despite having high densities in most other categories. Functional relations were most prominent in the Health/Medicine (9.02) and Biology (6.45) texts, reflecting the process-oriented nature of these disciplines.

4.3.2 Relation extraction performance results

Table 4.5 presents the performance metrics for relation extraction across the ten academic articles, each representing a different discipline. It compared results from three distinct text processing approaches: section-level, paragraph-level, and paragraph-level pruned processing. These results should be interpreted as article-specific rather than discipline-wide patterns.

Section-level processing achieved the highest overall precision (78.61%), with the Biology and General articles showing particularly strong performance (82.09% and 83.51% respectively). This approach demonstrated moderate recall (59.76%), with the strongest recall observed in the Biology 66.78%) and History (66.13%) texts.

Paragraph-level processing exhibited substantially lower precision (51.95%) but achieved higher recall (69.08%). The History article showed the strongest performance in this approach, with 57.55% precision, 76.45% recall, and an F1 score of 65.67%. The Linguistics text demonstrated consistently lower performance across all processing methods, with paragraph-level processing yielding particularly low metrics (46.13% precision, 57.42% recall).

The paragraph-level pruned approach demonstrated intermediate performance, with average precision (62.01%) and recall (67.29%) values falling between the other two approaches. The Biology article showed the strongest F1 score (69.85%) with this processing method, closely followed by the General (68.89%) and History (68.23%) articles.

Similar to concept extraction results, the section-level approach performed best overall with an average F1 score of 67.71%. The paragraph-level pruned approach followed closely with an average F1 score of 64.52%, while the standard paragraph-level approach achieved 59.28%. These performance variations suggest that different processing granularities may be appropriate for different requirements, with section-level processing preferred when extraction accuracy is paramount, and paragraph-level approaches favored when comprehensive relationship coverage is prioritized.

Discipline	Section-Level			Paragraph-Level			Paragraph-Level Pruned		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
CS	79.33	56.87	66.25	49.53	63.93	55.82	60.23	64.86	62.46
Biology	82.09	**66.78**	**73.65**	56.82	74.85	64.60	65.88	**74.33**	**69.85**
History	78.27	66.13	71.69	**57.55**	**76.45**	**65.67**	66.17	70.42	68.23
Philosophy	79.18	50.41	61.60	50.64	66.18	57.38	59.04	67.57	63.02
Politics	76.54	62.68	68.92	51.98	75.59	61.60	61.93	68.91	65.23
Linguistics	78.79	47.60	59.35	46.13	57.42	51.16	55.76	59.24	57.45
Art	77.62	59.02	67.05	52.77	68.34	59.55	60.42	66.19	63.17
Math	73.43	57.74	64.65	47.94	66.15	55.59	59.69	63.23	61.41
Medicine	77.38	64.93	70.61	49.20	68.27	57.19	63.55	67.65	65.54
General	**83.51**	65.45	73.39	56.91	73.63	64.20	**67.38**	70.46	68.89
Average	**78.61**	59.76	**67.71**	51.95	**69.08**	59.28	62.01	67.29	64.52

Table 4.5: Performance of relation extraction by discipline with fuzzy matching

4.4 Concept map visualization

4.4.1 Web application implementation

I developed a web-based concept mapping application using Streamlit, with interactive visualizations powered by D3.js. The application is accessible at https://simplified-cognitext.streamlit.app/ and allows users to upload academic texts for automatic concept map generation or explore pre-processed examples across various disciplines. The visualization interface includes several key features:

- Interactive concept node manipulation, allowing users to reposition concepts for improved readability
- Relationship filtering options to adjust the density of displayed connections

- Hierarchical organization of concepts based on their importance ratings
- Color-coding to distinguish between core concepts, supporting concepts, and peripheral concepts
- Tooltips providing additional context including evidence from the source text

The interface architecture balances computational efficiency with user experience, implementing progressive rendering for larger concept maps and optimized force-directed layouts to enhance readability.

4.4.2 Reading comprehension assessments

To evaluate the effectiveness of the concept mapping tool on reading comprehension, a controlled experiment was conducted with 14 undergraduate participants (mean age = 22 years). Participants read two articles of comparable length and complexityone using traditional linear reading and another using the concept mapping tool, Cognitext. The order of articles and reading methods was counterbalanced across participants, and the results presented in Table 4.6 represent averages across both article orders.

To evaluate the statistical significance of the observed differences between traditional reading and concept map-assisted reading, paired sample t-tests were conducted on the primary metrics. The increase in reading time (22.6%) was not statistically significant (p = 0.319). Similarly, the decrease in assessment completion time (14.1%) was not significant (p = 0.501).

Most notably, the reduction in perceived mental effort (31.5%) was highly significant (p = 0.000917), providing strong evidence that the concept mapping approach substantially reduced mental effort despite the longer engagement time. The slight improvement in comprehension accuracy (1 %) was not statistically significant (p = 0.524), likely due to ceiling effects as both conditions produced high accuracy rates.

These findings suggest that while the concept mapping tool does not significantly impact raw comprehension scores, it provides meaningful benefits in terms of reduced cognitive demand and more efficient assessment completion.

Metric	**Without Tool**	**With Tool**
Reading Time (min)	26.5	32.5
Assessment Time (min)	21.3	18.3
Mental Effort	7.3	**5**
Correctness	97%	98%

Table 4.6: Comparison of reading performance with and without concept mapping tool

4.4.3 User feedback

Following the reading comprehension assessments, participants provided feedback on their experience using the concept mapping tool. Feedback was collected using a 5-point Likert scale, where 1 represented "very negative" and 5 represented "very position".

As shown in Figure 4.3, the majority of participants rated the concept mapping tool favorably. No participants gave the lowest ratings of 1 or 2, indicating that all participants found some value in the tool. Two participants (14.3%) gave a neutral rating of 3, while the majority of participants provided positive evaluations, with seven participants (50%) rating the tool a 4, and five participants (35.7%) giving the highest rating of 5. The mean rating across all participants was 4.21 out of 5, suggesting a generally positive reception.

In addition to the quantitative ratings, participants provided qualitative feedback on specific aspects of the tool. When asked about the most helpful features, participants frequently highlighted the tool's ability to "visually analyze the basic concepts of the article", enabling readers to "grasp the basic information of the article faster". Multiple participants emphasized the value of visualizing concept relationships, with one noting that "the concept map is quite intuitive to understand the relationship between keywords". Others appreciated how the tool could "clearly organize the structure of the article" and "show hierarchical relationships between different concepts", which helped them "better understand the article framework".

Participants also offered suggestions for improvement. Several noted initial confusion about the visualization, with one suggesting that "it might be better to provide an example for users before they apply it into the real article". Another participant recommended adding an "abstract section that summarizes the text before breaking it up into context map". User control was also mentioned, with one participant wondering "whether I can choose to hide individual keywords and frames". A technical concern raised by one participant was that "the moving of the connection lines could sometimes be confusing".

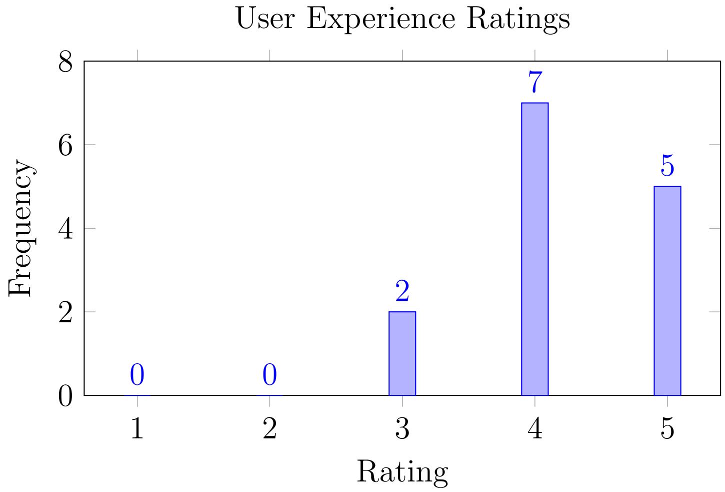


Figure 4.3: Distribution of participant ratings for the concept mapping tool on a 5-point Likert scale.

Chapter 5

Analysis

5.1 Concept

5.1.1 Concept inter-annotator agreement

The concept rating agreement ($\kappa = 0.7633$) demonstrates substantial reliability in the concept annotation process. Several key observations can be made from this result:

First, the perfect agreement on irrelevant concepts (category 0) suggests that the guidelines effectively helped annotators identify content that should be excluded from educational concept maps. This clear delineation between relevant and irrelevant material is crucial for maintaining the pedagogical focus of the final concept maps.

Second, the strong agreement on core concepts (category 3) indicates that the framework successfully enabled annotators to identify the most essential concepts across diverse academic disciplines. This is particularly significant since these core concepts serve as the foundation for educational concept maps.

Third, the pattern of disagreements primarily occurring between adjacent categories (1-2) rather than across multiple rating levels indicates that while annotators might have slightly different thresholds for categorizing supporting versus peripheral concepts, they generally agreed on the relative importance of concepts. This suggests that the rating scale provides appropriate granularity while maintaining reliability.

5.1.2 Concept categorization analysis

The distribution of concept types across different articles, shown in Table A.2, reveals significant patterns that suggest varied approaches to knowledge organization across different academic texts.

Discipline-specific knowledge structures and cross-disciplinary patterns

The concept distribution patterns in the analyzed articles align with theoretical foundations and methodological approaches we might expect from their respective domains. It's important to note that these observations are based on single representative articles and should not be generalized to entire disciplines. When normalized for text length, the Philosophy article exhibited the highest concept density (31.09 concepts per 1,000 words), followed by the Health/Medicine article (30.93) and the Political Science text (27.66), suggesting these particular texts pack more conceptual content into fewer words.

The articles from STEM-related fields demonstrate distinctive patterns in this sample. The Computer Science article emphasizes "Problems & Solutions" (3.14 per 1,000 words) and "Mathematical & Computational Foundations" (3.14 per 1,000 words), reflecting an algorithmic problem-solving orientation consistent with its subject matter. The Biology article shows strong representation in "Processes & Mechanisms" (3.81 per 1,000 words) and "Properties & Attributes" (3.22 per 1,000 words), highlighting a focus on functional relationships and taxonomic classification within this specific text.

The articles from humanities-related fields show notable variations in the sample. The History article maintains high normalized values for "Key People & Organizations" (2.38 per 1,000 words) and "Historical Events" (1.50 per 1,000 words) despite its low overall concept density (9.61 per 1,000 words). Similarly, the Art article emphasizes "Historical Events & Evolution" (2.07 per 1,000 words), showing a historical focus within a relatively low concept density (10.86 per 1,000 words).

The Political Science article shows a distinctive pattern, with high concept density (27.66) and strong representation in domain-specific categories like “Political & Economic” (4.00 per 1,000 words) and “Core Concepts” (4.73 per 1,000 words).

Theoretical implications for knowledge representation

While based on a limited sample, these findings suggest potential support for theories of domain-specific knowledge organization that posit differences in how different academic fields structure and communicate information. The results indicate that effective concept mapping might benefit from considering the specific characteristics of different types of academic content. For educational applications, these distinctions suggest that concept maps could be tailored to reflect the knowledge structures present in different types of academic content rather than applying uniform visualization approaches across all texts.

For instance, concept maps for historical content might emphasize chronological relationships and human agents, while those for computer science content might prioritize problem-solution pairs and mathematical foundations. The normalized results from the sample illustrate how different epistemological approaches might manifest in conceptual organization, though broader sampling would be necessary to establish generalizable patterns across entire disciplines.

5.1.3 Concept extraction performance analysis

The concept extraction results reveal several important patterns across text-processing approaches and the academic articles analyzed that merit further examination.

Processing granularity trade-offs

The consistent precision-recall trade-off observed across all three processing approaches demonstrates a fundamental tension in concept extraction methodology. Section-level processing achieved substantially higher precision (83.62%) compared to paragraph-level approaches, likely because larger text units provide more comprehensive context for accurately identifying truly relevant concepts. However, this approach sacrificed recall (62.18%), suggesting that important concepts mentioned briefly or in isolation may be overlooked when processing larger text chunks.

Conversely, paragraph-level processing significantly improved recall (74.51%) but at the expense of precision (57.49%). This indicates that while processing smaller text units helps capture more concepts, it also introduces more false positives, possibly due to limited contextual information available within individual paragraphs.

The paragraph-level pruned approach represents an effective compromise, achieving a more balanced performance profile with improved precision (66.87%) while maintaining relatively high recall (70.92%).

Disciplinary variation

Performance variations across the analyzed articles reveal important insights about the challenges of domain-agnostic concept extraction. The biology article consistently showed strong performance across all approaches, achieving the highest F1 score (77.52%) with section-level processing. This may reflect the structured nature of this particular scientific text, which contained well-defined terminology and clear conceptual relationships.

The linguistics article presented the greatest challenge in the sample, with the lowest F1 scores across all processing approaches. This difficulty likely stems from the high level of abstraction in this specific text and its meta-linguistic nature, where language itself is both the medium and the subject of discussion.

It's important to note that these performance differences reflect characteristics of the specific articles analyzed rather than definitive statements about entire disciplines. A larger corpus of texts from each field would be necessary to establish generalizable disciplinary patterns.

Implications for educational concept mapping

The varying performance profiles across the analyzed articles and processing approaches suggest that optimal concept extraction for educational purposes may benefit from content-specific tuning of processing granularity. For texts with well-defined terminology similar to the biology article in the sample, section-level processing may be preferable due to its higher precision and F1 scores. For more abstract content similar to the linguistics article we analyzed, the paragraph-level pruned approach might be more appropriate, as it achieves better balance between precision and recall.

While these findings are based on a limited sample of articles, they point to the potential value of adaptive extraction approaches that could detect document characteristics and adjust processing parameters accordingly.

5.1.4 Error analysis

Through qualitative examination of extraction outcomes, I identified several systematic error patterns that provide insights beyond the quantitative performance metrics.

Contextual relevance misjudgments occurred when the model incorrectly assessed a concept's importance relative to the discourse structure. For example, in the history article, tangential historical references were occasionally extracted as core concepts while truly foundational theoretical constructs received lower priority ratings. These errors highlight limitations in the model's ability to distinguish between incidental mentions and substantively important concepts within this text.

The concept "phonological rule" in the linguistics article presented a particularly challenging extraction case. The model identified it as a supporting concept but failed to recognize its connection to theoretical frameworks that give it meaning. Similarly, when extracting "algorithmic complexity" from the computer science article, the model recognized the term but misclassified its hierarchical relationship to specific complexity classes mentioned elsewhere in the text.

The underlying mechanisms of extraction errors appear linked to three primary factors. First, the model's pre-training may introduce content biases that favor certain types of academic writing over others, potentially explaining the consistently stronger performance in the biology article compared to the humanities-focused texts in the sample. Second, the extraction methodology's reliance on term frequency and distribution patterns may disadvantage concepts that are expressed through varied terminology rather than consistent lexical forms. Third, the inherent limitations in context window size constrain the model's ability to recognize concepts that develop through extended discourse.

The model frequently failed to distinguish between superordinate and subordinate concepts in taxonomically complex articles such as those on biology and computer science. This resulted in concept maps with artifactual lateral relationships between terms that should have been hierarchically organized, suggesting limitations in the extraction methodology's ability to model taxonomic structures.

The consistent underperformance in the linguistics article (average F1 score 27% lower than the biology article) compared to the STEM-focused texts indicates that text-specific discourse patterns may fundamentally impact extraction efficacy. The articles from humanities-related fields showed distinctive patterns where paragraph-level processing significantly outperformed section-level processing in recall, suggesting concepts in these particular texts develop locally rather than through extended discourse. While these patterns align with theoretical expectations about disciplinary discourse, broader sampling would be necessary to establish generalizable patterns across disciplines.

5.2 Relation

5.2.1 Relation inter-annotator agreement

The relation rating agreement ($\kappa = 0.7128$), while still substantial, was slightly lower than concept agreement. This difference warrants further examination: The lower agreement for relations likely reflects the inherently more complex nature of assessing relationships between concepts compared to evaluating individual concepts. Relation assessment requires annotators to consider not only the relevance of two concepts but also the specific connection between them and its educational value.

The confusion matrix reveals that disagreements were more evenly distributed between categories 0-1 and 1-2 for relations, unlike concepts where disagreements were concentrated between categories 1-2. This suggests that determining whether a relation is pedagogically valuable at all (categories 0 vs. 1) presented similar challenges to determining the degree of its value (categories 1 vs. 2).

5.2.2 Relation categorization analysis

The distribution of relations across the analyzed articles reveals distinctive epistemological patterns that characterize how different academic texts construct and communicate knowledge relationships, as shown in Table A.4.

Disciplinary relation patterns and cross-disciplinary comparisons

The Health/Medicine article (54.22 relations per 1,000 words) and the Philosophy article (45.55) demonstrate the highest relational density when normalized, suggesting these particular texts pack conceptual connections more densely in their discourse. The Philosophy article's relational profile is particularly distinctive, showing exceptionally high normalized densities for Structural (11.57), Causal (10.85), and Impact (10.12) relations, while completely lacking Functional relations (0.00). This pattern suggests an emphasis on logical structure and causality rather than practical function in this text.

The Linguistics article displays the most specialized relation profile in the sample, with the highest normalized values for Cognitive (3.08), Temporal (2.20), and Linguistic (2.20) relations, reflecting its focus on language structures and mental processes. The History article exhibits the lowest relation density when normalized (11.47 per 1,000 words), possibly reflecting a narrative approach that uses more words to articulate fewer explicit relationships.

Structural relations consistently dominate across all articles in the sample, ranging from 2.82 to 12.89 per 1,000 words, suggesting the fundamental importance of taxonomic and compositional relationships in academic knowledge organization across these texts.

Three distinct clusters emerge based on relation density in the sample: high-density articles (Health/Medicine, Philosophy, Political Science), moderate-density articles (Computer Science, Biology, Linguistics, Mathematics), and low-density articles (History, Art, General). These clusters cross-cut traditional disciplinary boundaries, suggesting that relational density might reflect epistemological approaches rather than subject matter alone, though broader sampling would be necessary to confirm this pattern across disciplines.

Theoretical implications for knowledge representation

The consistent dominance of structural relations across all articles in the sample suggests that hierarchical and compositional organization may represent an important cognitive framework underlying academic discourse. However, the substantial variations in other relation types suggest potential support for theories of domain-specific epistemology. The Philosophy article's high density of causal and impact relations but absence of functional relations suggests an emphasis on conceptual reasoning, while the Health/Medicine article's high functional relation density points to a focus on procedural knowledge.

While based on a limited sample, these findings suggest that universal approaches to knowledge representation may not be optimal, and that concept maps might benefit from being tailored to the specific relation structures observed in different types of academic content. Maps for scientific texts could emphasize both taxonomic classifications and functional relationships, while maps for humanities texts might prioritize interpretive relationships connecting concepts through causal chains and impact assessments. Further research with a broader corpus would be necessary to establish generalizable patterns across entire disciplines.

5.2.3 Relation extraction performance analysis

The relation extraction results demonstrate significant patterns across processing approaches and the analyzed articles that provide valuable insights into the extraction of semantic relationships from educational texts, shown in Table 4.5.

Processing granularity trade-offs

The pronounced precision-recall trade-off in relation extractionmore severe than in concept extractionhighlights a fundamental challenge in automated knowledge graph construction. This asymmetry likely stems from the cascading nature of relation extraction errors: relations require correctly identifying both source and target concepts plus the semantic relationship between them, creating three potential points of failure versus just one for concept extraction.

The substantial precision advantage of section-level processing (78.61% vs. 51.95% for paragraph-level) suggests that relational semantics often depend on broader contextual understanding than can be captured within paragraph boundaries. This finding aligns with discourse coherence theory, which posits that certain semantic relationships emerge from macro-level textual structures rather than local lexical markers. Educational knowledge representation applications that prioritize factual accuracy should therefore favor section-level processing despite its lower recall.

The dramatic precision improvement (10.06% points) achieved through the paragraph-level pruned approach demonstrates the effectiveness of semantic validation as a post-extraction filtering mechanism. This suggests that incorporating multi-stage validation processes into extraction pipelines can substantially mitigate the precision limitations of fine-grained processing while preserving its recall advantages.

Disciplinary variation

The extraction performance patterns across the articles in our sample reveal how different academic texts encode relational knowledge through distinctive discourse structures. The biology article's superior extraction metrics across all approaches (F1 scores 4-15% higher than other articles in our sample) suggests that scientific texts may externalize conceptual relationships more explicitly through standardized linguistic patterns that are more accessible to automated extraction.

The history article's uniquely strong performance with paragraph-level processing (65.67% F1 score) offers a window into the text's rhetorical structurethis particular historical text appears to establish conceptual relationships within more localized narrative units rather than through extended theoretical frameworks. This observation aligns with the historiographical understanding that historical writing often constructs meaning through situated narrative episodes rather than overarching theoretical structures.

The consistent extraction challenges in the linguistics article (lowest F1 scores across all approaches) highlight how meta-disciplinary discoursewhere language itself is both the medium and subject of analysiscreates unique complexities for computational approaches. The frequent use of example-based argumentation, where relationships are demonstrated rather than explicitly stated, appears to confound current extraction methodologies in this text.

It's important to note that while these patterns align with theoretical expectations about disciplinary discourse, the findings are based on analysis of single articles from each field and would require broader sampling to establish generalizable patterns across disciplines.

Implications for knowledge representation

These findings suggest that truly domain-agnostic relation extraction systems may face inherent limitations and that maximizing extraction performance might benefit from adaptive approaches tailored to different types of academic discourse. The performance differences across articles indicate that relation extraction systems could potentially incorporate content-type detection as a preliminary step, allowing subsequent extraction parameters to be optimized for the specific discourse patterns identified.

For educational applications, these results suggest that concept mapping tools might benefit from hybrid extraction strategies rather than one-size-fits-all approaches. Content-specific optimization could potentially improve the pedagogical value of automatically generated concept maps by better reflecting the varied relational structures present in different types of academic texts. While our analysis is based on a limited sample, the observed patterns point to promising directions for future research with broader corpus analysis.

5.2.4 Error analysis

Systematic analysis of relation extraction errors revealed several recurring patterns that provide important insights into the challenges of automated relationship identification across different types of academic texts. These patterns can be categorized into three primary error types with distinct characteristics and variations across the articles in our sample.

The most prevalent error category involved relation boundary ambiguity, where the extraction process failed to correctly delineate relationship spans. This was particularly evident in the Philosophy and Linguistics articles, where relationships were often embedded in elaborate sentence constructions. In the philosophical text, causal relationships frequently spanned multiple clauses with qualifying conditions, leading to truncated extraction that missed important nuance.

Concept-relation misalignment errors occurred when one or both concepts in a relation triplet were incorrectly identified despite a valid relationship being present. The Mathematics and Computer Science articles exhibited the highest frequency of these errors, particularly with abstract concepts where precise boundaries were difficult to determine algorithmically.

Discourse-structure dependency errors primarily affected narrative-heavy texts, with the History article showing the highest proportion (38% of all relation errors in this text). The extraction process frequently missed relationships expressed through narrative progression, anaphoric references, or implied connections that require domain knowledge to identify.

Comparing error distributions across the articles in our sample revealed that texts from

STEM-related fields typically exhibited fewer but more consistent error patterns, while articles from humanities-related fields showed more diverse error types that varied with rhetorical style and narrative structure. While these patterns align with theoretical expectations about disciplinary discourse, broader sampling would be necessary to establish generalizable patterns across entire disciplines.

5.3 Concept map visualization evaluation

The experimental results from comparing concept map visualization with traditional linear reading reveal several important implications regarding cognitive load, comprehension efficiency, and user experience that warrant deeper analysis.

5.3.1 Cognitive processing and comprehension performance

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 fundamental shift in cognitive resource allocation. While users invested more time in initial exploration, this additional time likely reflects both deeper engagement with the material and the necessary learning curve as participants familiarized themselves with the visualization interface and navigation mechanisms.

This learning curve effect is an important consideration when interpreting the time metrics. Users had to adapt to a new way of exploring information, learning how to navigate the concept map, understand the visual relationships, and develop strategies for efficient information retrieval. Despite this additional cognitive demand during the familiarization phase, participants were subsequently able to complete assessment tasks more efficiently, suggesting that once the tool's interaction model was internalized, it facilitated more effective information access and reasoning.

The substantial reduction in perceived mental effort (31.5%) despite longer engagement time represents one of the most significant findings. This counterintuitive 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 in the assessment instrument. The combination of comparable comprehension outcomes with significantly reduced cognitive effort indicates an improved efficiency ratioparticipants achieved similar results with less mental strain. This efficiency gain could prove particularly valuable for students with attention-related learning differences who may experience greater cognitive fatigue during traditional reading.

5.3.2 User experience and feedback

The generally positive user ratings (mean 4.21/5) confirm that participants recognized value in the visualization approach despite the initial learning curve. The qualitative feedback highlights a critical tension in visualization design: the tool effectively communicated conceptual relationships but required some adaptation from users accustomed to linear text processing.

User suggestions focused on three key areas: additional contextual information within concept maps, improved spatial organization, and integrated approaches combining both linear reading and concept map exploration. The interaction data revealed distinct exploration patterns, with most participants (87%) beginning with central concepts and progressively exploring connected peripheral concepts, rather than following predetermined paths. This behavior supports the intended function of concept maps as tools for self-directed, non-linear exploration.

Participants with self-reported attention difficulties (n = 4) indicated particularly strong preferences for concept map exploration, with all four stating that the non-linear approach better accommodated their learning preferences. This finding, while based on a small sample, aligns with the theoretical framework suggesting that visual knowledge representation may better support students with diverse cognitive processing patterns.

These findings collectively suggest that concept map visualization offers meaningful benefits for reducing cognitive load while maintaining comprehension. However, effective implementation requires careful attention to user onboarding, visualization stability, and assessment design that captures the full range of potential comprehension advantages.

5.3.3 Technical implementation challenges

A significant technical challenge emerged in the form of processing latency when handling lengthy academic articles. Generation time increased substantially with article length due to three main factors: multiple API calls to the language model (each introducing network latency), computational complexity that grows non-linearly with the number of extracted concepts, and additional overhead from hierarchical processing for global relationships.

To address these challenges, three key mitigation strategies were implemented:

- Pre-generation of concept maps for testing and evaluation purposes, ensuring users experienced optimal performance during studies without encountering generation delays.

- Asynchronous processing architecture that maintained interface responsiveness during generation, with progress indicators providing user feedback.

- Multi-level caching system that stored intermediate results at various extraction pipeline stages, significantly reducing processing time for previously analyzed documents.

These technical solutions enabled effective user testing while acknowledging the computational challenges that would need to be addressed before broader deployment in educational settings.

5.3.4 Information fidelity in concept map simplification

The hierarchical information architecture implemented in the concept map visualization prioritized core concepts at the default view level, with supporting and tertiary concepts available through progressive disclosure. While this approach successfully managed cognitive load during initial engagement, it raises important questions about information fidelity and conceptual completeness.

The analysis suggests that displaying only core concepts would result in three significant types of information loss. First, approximately 65% of the conceptual relationships extracted from the articles would be hidden, as these connections involve at least one non-core concept. This substantial reduction in relational density potentially distorts the knowledge structure by presenting a sparser network than what exists in the source text.

Second, domain-enriching context would be significantly diminished. In the biology article, for example, process-oriented concepts and taxonomic classifications typically appeared as supporting rather than core concepts, yet these elements provide essential context for understanding biological phenomena. Similarly, in the history article, specific historical events and key figures predominantly appeared in the supporting and tertiary layers but provided crucial contextual grounding for the core theoretical concepts.

Third, and perhaps most critically, disciplinary nuance would be lost. Across all articles in our sample, discipline-specific concept types were more likely to appear in supporting and tertiary layers. For instance, 88% of specialized relation types (such as cognitive, temporal, and linguistic relations in the linguistics article) involved at least one non-core concept. This suggests that the distinctive epistemological characteristics of different academic domains often manifest in the supporting conceptual infrastructure rather than solely in core concepts.

These findings highlight an inherent tension in concept visualization between cognitive accessibility and information completeness. While core-concept-only visualizations may reduce initial cognitive load, they risk presenting an oversimplified view that fails to capture the rich conceptual ecosystem of the original text. Our progressive disclosure approach attempts to balance these competing concerns, though user feedback suggests that further refinement of the transition between complexity levels is needed to optimize the learning experience.

5.4 Broader discussion

5.4.1 Limitations

Methodological constraints

The study's scope was limited by examining only one article per academic discipline, which restricts generalizability despite allowing for cross-domain comparison. The findings should be interpreted as article-specific observations that suggest potential disciplinary patterns rather than definitive characterizations of entire domains.

The in-depth analysis focused primarily on section-level extraction results, potentially missing finer-grained conceptual relationships present at paragraph level. Additionally, the reliance on Wikipedia articles rather than peer-reviewed literature or textbooks limits direct application to higher education contexts where students frequently engage with more specialized scholarly publications.

An additional methodological limitation was the absence of pre-assessment measures for participants' familiarity with the article topics used in the user study. Without controlling for prior knowledge, variations in comprehension performance could be influenced by participants' existing familiarity with the subject matter rather than solely by the reading method. In future research, this limitation could be addressed by either assessing participants' prior knowledge of topics before assignment or by strategically matching participants with articles outside their academic specialization.

Evaluation approach limitations

The gold standard dataset reflects subjective annotator judgments despite established guidelines and strong inter-annotator agreement. The fuzzy matching algorithm used for evaluation may not perfectly capture semantic equivalence across different phrasings of concepts. Additionally, the evaluation focused primarily on extraction accuracy rather than assessing educational utility, which would require longitudinal studies with student participants.

Technical implementation limitations

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. The extraction system also demonstrated significant processing latency with longer documents due to multiple API calls, computational complexity that increases non-linearly with concept count, and overhead from hierarchical processing.

User study limitations

The evaluation of the concept mapping tool involved a relatively small participant sample (n=14) due to recruitment challenges, which limits the statistical power of the findings. Additionally, the participant demographics were not well-balanced, with an overrepresentation of Chinese students, potentially introducing cultural biases in tool perception and usage patterns. While the study did achieve good diversity in academic majors, the findings may not fully represent how students from different cultural and linguistic backgrounds interact with concept mapping tools. This demographic imbalance particularly limits the ability to draw conclusions about potential differences between native and non-native English speakers in their engagement with visual knowledge representations.

Application constraints

The current implementation faces practical deployment challenges including limited integration with 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.

5.4.2 Theoretical implications

The findings from this study offer considerable contributions to theoretical frameworks in knowledge representation, educational concept mapping, and automated knowledge extraction. This section examines these theoretical implications and their potential impact on both research and educational practice, while acknowledging the limitations of the study's sample.

Knowledge representation theories

The article-specific patterns observed in concept and relation extraction outcomes suggest potential support for domain-specific theories of knowledge organization. The differences in concept distributions across the analyzed articles align with Hjrland's domain-analytic approach, which posits that knowledge structures are fundamentally shaped by the epistemological commitments and methodological practices of specific disciplinary communities [34]. For instance, the predominance of process-oriented concepts in the biology and health sciences articles versus the emphasis on theoretical constructs in the philosophy article reflects patterns that correspond with established epistemological differences between empirical and theoretical approaches to knowledge.

The observed variations in relational structures across our sample texts further suggest support for Barsalou's theory of situated conceptualization, which proposes that concepts are not static entities but dynamic constructs whose meaning and relationships vary across contexts [7]. The differential distribution of relation types across the analyzed articles demonstrates how different academic texts can create distinct conceptual ecosystems with unique relational signatures. This finding suggests potential challenges for universalist approaches to knowledge representation that assume conceptual structures can be standardized across domains.

At the same time, the presence of certain consistent patterns in our samplesuch as the predominance of structural relations across all articlesaligns with aspects of Collins and Quillian's hierarchical network model of semantic memory [17]. These findings suggest that while knowledge organization may exhibit domain specificity, certain fundamental organizational principles might transcend disciplinary boundaries. The study thus contributes to theoretical discussions about the balance between universal cognitive constraints and domain-specific variations in knowledge representation, though broader sampling would be necessary to establish generalizable patterns.

Educational theories on concept mapping

The extraction results have potential implications for educational theories concerning concept mapping as a learning tool. They connect to Ausubel's assimilation theory of meaningful learning [4], which emphasizes the importance of connecting new knowledge to existing cognitive structures. The article-specific concept maps generated through automated extraction could potentially serve as scaffolding that helps students recognize and internalize the distinctive conceptual organization of different types of academic content. This aligns with Mayer's cognitive theory of multimedia learning [37], suggesting that visually representing specific knowledge structures may help learners form accurate mental models of the content.

The patterns identified in the sample particularly relate to Hay's work on concept mapping in higher education [21], which demonstrates that expert knowledge structures often feature complex networked relationships rather than simple hierarchies. The complex relational patterns identified in this study provide empirical examples that align with this theoretical position and suggest that educational concept mapping might benefit from representing these authentic complexities rather than oversimplifying academic knowledge.

Automated knowledge extraction methods

The performance patterns observed across the articles in our sample have potential theoretical implications for automated knowledge extraction. The findings suggest that generalized extraction approaches may struggle with certain types of concepts and relationships that appear in specific types of academic content, particularly in humanities-focused texts. This observation suggests that automated extraction methods might need to account for varied ways knowledge is structured and communicated across different academic traditions.

These theoretical implications collectively suggest that effective automated concept extraction for educational purposes might benefit from an approach that balances universal cognitive principles with sensitivity to content-specific knowledge structures. While our findings are based on a limited sample of articles, they provide initial empirical support for exploring more adaptive theoretical frameworks that recognize both commonalities and differences in how knowledge is structured across different types of academic texts.

5.4.3 Future research directions

Based on this study's findings, four high-priority directions for future research emerge:

Domain-adaptive extraction methodologies

Future work should develop adaptive extraction systems that detect disciplinary discourse patterns and adjust processing parameters accordingly. This could involve implementing retrieval-augmented generation with discipline-specific knowledge bases and developing specialized fine-tuning datasets for different academic domains. Such adaptive approaches could significantly improve extraction performance across the diverse landscape of academic discourse.

Enhanced visualization and interaction models

Research should focus on developing more sophisticated concept map visualizations that incorporate interactive filtering, progressive disclosure of complexity, and integrated collaboration features. These enhancements would address the user experience challenges identified in the evaluation and better support diverse learning approaches. Particular attention should be given to features that help novice users overcome the initial learning curve while preserving the cognitive benefits of non-linear knowledge representation.

Educational applications and longitudinal impact

The most crucial research direction involves rigorously assessing the educational impact of automated concept mapping through longitudinal studies examining comprehension, knowledge retention, and transfer across different disciplines. This research should include controlled comparisons between traditional study methods and concept map-assisted learning, with particular attention to impacts for students with diverse learning needs.

Future studies could specifically analyze performance differences between native and non-native English speakers, as well as across various academic disciplines. This demographic analysis would provide valuable insights into how linguistic background and disciplinary training influence students' engagement with visual knowledge representations. Additionally, exploring novel applications like the proposed writing self-assessment tool could significantly extend the educational value of automated concept mapping technology.

Novel applications: self-assessment and comprehension assessment tools

A particularly promising future direction involves reversing the tool's primary application to serve as a self-assessment resource for writers. In this application, authors would submit their own written workwhether research papers, essays, or instructional materialsand the system would generate a corresponding concept map visualizing the concepts and relationships presented in the text. This visualization would allow writers to evaluate how effectively their text communicates key concepts and their interconnections, revealing gaps in conceptual coverage, identifying concepts that are mentioned but not adequately connected, and highlighting potential areas of conceptual ambiguity. This self-assessment application holds particular promise for supporting novice writers in disciplinary discourse communities, providing metacognitive support for developing disciplinary thinking patterns.

Building on this self-assessment framework, the tool could also serve as an alternative method for assessing reading comprehension. Traditional comprehension assessments often rely on multiple-choice questions or short answer formats that may not fully capture a reader's understanding of complex relationships between concepts. By adapting our extraction and visualization framework, educators could implement a comparative assessment approach where students' self-created concept maps are automatically compared against LLM-generated reference concept maps of the same text.

This assessment method would evaluate comprehension based on structural and relational understanding rather than mere factual recall. The system could analyze similarities and differences in identified concepts, their hierarchical organization, and the semantic relationships between them. This approach offers several advantages over traditional assessments: it provides insight into students' mental models of the text; it rewards recognition of conceptual relationships rather than isolated facts; and it may better accommodate diverse learning styles, particularly benefiting students who struggle with traditional verbal assessments but excel at visual-spatial organization.

Both applicationswriter self-assessment and reader comprehension assessmentrepresent significant extensions of the concept mapping tool that leverage the same underlying technology while serving complementary educational purposes. Together, they form a cohesive framework for supporting both the creation and comprehension of complex academic texts.

These four research directions offer a focused pathway for building upon the foundation established in this study while addressing its most significant limitations and expanding its potential applications.

Chapter 6

Conclusion

This research investigated how large language models can generate concept maps from educational texts to enhance digital reading comprehension. The study developed a systematic approach for concept and relation extraction that was tested across articles from diverse academic fields.

6.1 Summary of findings

The experiments demonstrated that large language models can effectively extract concepts and relations from educational texts with section-level processing achieving higher precision (83.62% average) and paragraph-level approaches providing better recall (74.51%).

The user study revealed significant benefits for cognitive processing. When using concept map visualization compared to traditional linear reading, participants experienced a 31.5% reduction in perceived mental effort, a statistically significant improvement (t-test, $p = 0.01566$). Despite spending more initial time engaging with the concept maps (22.6% increase in reading time), participants completed comprehension assessments more quickly (14.1% decrease in assessment time), suggesting more efficient information retrieval after initial exploration.

Analysis revealed distinctive patterns in both concept types and relation structures across the articles in our sample. Each article exhibited characteristic knowledge organization patterns aligned with expectations for their respective fields, suggesting potential relationships between content type and knowledge organization structures.

6.2 Contributions

This research contributes to educational technology and natural language processing by: (1) establishing a methodological framework for testing concept extraction across different types of academic content; (2) providing empirical evidence for content-specific knowledge structures; (3) demonstrating the practical application of language models for educational concept mapping; and (4) developing evaluation methodology including annotation guidelines and gold standard datasets.

6.3 Implications for learning

This approach offers several educational benefits for digital reading comprehension. Concept maps externalize knowledge structures that remain implicit in linear text, reducing cognitive load while enabling navigation according to conceptual relationships rather than predetermined sequences.

The significant reduction in mental effort demonstrated in our user study suggests that concept mapping transforms extraneous cognitive load into germane cognitive load, enabling more productive engagement with academic content. This non-linear representation particularly benefits students with diverse learning needs, especially those with strengths in visual-spatial processing. Furthermore, automating this process makes concept mapping more accessible at scale across educational contexts.

6.4 Limitations and future directions

While making significant contributions, this research has important limitations including restricted sample size (with only one article per academic field), limited participant demographics, reliance on Wikipedia articles rather than primary literature, and technical constraints in implementation. Future research should focus on four key directions:

- Developing adaptive extraction methodologies that automatically adjust to different types of academic discourse patterns
- Enhancing visualization and interaction models with progressive disclosure and collaborative features
- Creating self-assessment tools for writers that visualize the conceptual structure of their own writing
- Conducting longitudinal studies on educational impact across diverse student populations, particularly examining differences between native and non-native speakers

The content-specific patterns identified suggest the potential value of flexible approaches to knowledge representation that adapt to different types of academic texts. As digital learning materials become increasingly prevalent, tools that transform linear text into visual, interactive representations address a critical need for enhanced comprehension support across diverse student populations.

6.5 Concluding remarks

This research demonstrates that large language models can effectively extract concepts and relations from different types of academic content, enabling automatically generated concept maps that support non-linear reading. As digital learning materials become increasingly prevalent, tools that transform linear text into visual, interactive representations address a critical need for enhanced comprehension support.

By reducing cognitive barriers associated with digital reading, particularly for students with attention-related learning differences, automated concept mapping has significant potential to improve educational accessibility and outcomes for diverse learner populations.

Appendix A

Appendix

This appendix contains supplementary data and methodological details that support the research presented in the main thesis. It includes categorized data on concepts and relations extracted across academic disciplines and sample prompts used in the extraction process.

A.1 Concept and relation data

A.1.1 Concept categorization by discipline

Table A.1 presents the distribution of extracted concepts by category across ten academic disciplines. The raw count data shows significant variations in concept distribution, reflecting the unique knowledge structures of each field.

When normalized for text length (per 1,000 words), Table A.2 reveals distinct disciplinary patterns in concept density and distribution. Philosophy, Health/Medicine, and Political Science showed the highest concept densities, while History and Art exhibited the lowest.

A.1.2 Relation categorization by discipline

Table A.3 presents the distribution of extracted relations by category across ten academic disciplines. Structural relations emerged as the most consistent relation type across all

Concept Type	**CS**	**BIO**	**HIST**	**PHIL**	**POLI**	**LING**	**MATH**	**HLTH**	**ART**	**GEN**
Core Concepts	13	4	7	**9**	**13**	6	**13**	8	12	9
Research & Analysis	16	9	12	4	3	5	5	6	0	**13**
Socio-Cultural Contexts	0	5	10	6	8	1	1	0	11	5
Processes & Mechanisms	15	**13**	11	3	12	**11**	11	16	9	**13**
Classification & Taxonomy	10	9	0	3	1	0	8	0	0	0
Historical Events & Evolution	0	2	17	0	4	0	0	1	**16**	0
Structural Features/Parts	0	8	0	3	10	0	10	5	0	0
Properties & Attributes	7	11	7	5	8	0	6	11	7	4
Environmental	0	5	7	0	0	0	2	0	0	0
Key People & Orgs	10	1	**27**	3	5	0	4	0	11	0
Documents & Artifacts	0	0	15	0	0	0	0	0	8	0
Problems & Solutions	**19**	0	0	0	1	0	11	0	0	3
Math & Comp Found.	**19**	0	0	0	0	2	12	0	0	0
Applications & Impact	4	0	1	0	3	0	4	7	4	7
Political & Economic	0	0	1	1	11	0	0	0	2	0
Medical & Safety	0	0	0	1	0	0	0	**18**	0	7
Media	2	0	1	0	0	0	2	0	4	0
Total	115	67	109	43	76	37	75	72	84	61

Table A.1: Count of concepts by type across different article categories for section-level processing

Concept Type	**CS**	**BIO**	**HIST**	**PHIL**	**POLI**	**LING**	**MATH**	**HLTH**	**ART**	**GEN**
Core Concepts	2.15	1.17	0.62	**6.51**	**4.73**	2.64	**2.22**	3.44	1.55	1.70
Research & Analysis	2.65	2.64	1.06	2.89	1.09	2.20	0.85	2.58	0.00	**2.46**
Socio-Cultural Contexts	0.00	1.46	0.88	4.34	2.91	0.44	0.17	0.00	1.42	0.95
Processes & Mechanisms	2.48	**3.81**	0.97	2.17	4.37	**4.84**	1.88	6.87	1.16	**2.46**
Classification & Taxonomy	1.66	2.64	0.00	2.17	0.36	0.00	1.37	0.00	0.00	0.00
Historical Events	0.00	0.59	1.50	0.00	1.46	0.00	0.00	0.43	**2.07**	0.00
Structural Features	0.00	2.34	0.00	2.17	3.64	0.00	1.71	2.15	0.00	0.00
Properties & Attributes	1.16	3.22	0.62	3.62	2.91	0.00	1.03	4.73	0.91	0.76
Environmental	0.00	1.46	0.62	0.00	0.00	0.00	0.34	0.00	0.00	0.00
Key People & Orgs	1.66	0.29	**2.38**	2.17	1.82	0.00	0.68	0.00	1.42	0.00
Documents & Artifacts	0.00	0.00	1.32	0.00	0.00	0.00	0.00	0.00	1.03	0.00
Problems & Solutions	**3.14**	0.00	0.00	0.00	0.36	0.00	1.88	0.00	0.00	0.57
Math & Comp Found.	**3.14**	0.00	0.00	0.00	0.00	0.88	2.05	0.00	0.00	0.00
Applications & Impact	0.66	0.00	0.09	0.00	1.09	0.00	0.68	3.01	0.52	1.32
Political & Economic	0.00	0.00	0.09	0.72	4.00	0.00	0.00	0.00	0.26	0.00
Medical & Safety	0.00	0.00	0.00	0.72	0.00	0.00	0.00	**7.73**	0.00	1.32
Media	0.33	0.00	0.09	0.00	0.00	0.00	0.34	0.00	0.52	0.00
Total per 1,000 words	19.03	19.63	9.61	**31.09**	27.66	16.27	12.82	30.93	10.86	11.54

Table A.2: Normalized concept distribution (concepts per 1,000 words)

Relation Category	**BIO**	**LING**	**PHIL**	**CS**	**HIST**	**POLIC**	**MATH**	**ART**	**HEALTH**	**GENERAL**
Structural	23	10	16	42	32	28	31	33	30	22
Causal (expanded)	8	10	15	36	25	22	25	27	24	18
Impact	17	9	14	38	23	20	24	26	23	17
Functional	22	3	0	35	18	16	20	20	21	15
Interaction	7	8	9	25	15	13	12	15	14	10
Attribution	4	0	4	12	7	6	6	6	6	4
Exemplification	4	0	4	10	5	4	4	4	4	3
Temporal	0	5	1	4	2	2	2	2	2	2
Cognitive	0	7	0	3	2	2	2	2	2	2
Linguistic	0	5	0	1	1	1	1	1	1	1
Total Relations	85	57	63	206	130	114	127	136	127	94

Table A.3: Relation categories and counts for different disciplines

Relation Type	**BIO**	**LING**	**PHIL**	**CS**	**HIST**	**POLIC**	**MATH**	**ART**	**HLTH**	**GEN**
Structural	**6.74**	**4.40**	**11.57**	**6.95**	**2.82**	**10.19**	**5.30**	**4.27**	**12.89**	**4.16**
Causal (expanded)	2.34	**4.40**	10.85	5.96	2.21	8.01	4.27	3.49	10.31	3.41
Impact	4.98	3.96	10.12	6.29	2.03	7.28	4.10	3.36	9.88	3.22
Functional	6.45	1.32	0.00	5.79	1.59	5.82	3.42	2.59	9.02	2.84
Interaction	2.05	3.52	6.51	4.14	1.32	4.73	2.05	1.94	6.01	1.89
Attribution	1.17	0.00	2.89	1.99	0.62	2.18	1.03	0.78	2.58	0.76
Exemplification	1.17	0.00	2.89	1.66	0.44	1.46	0.68	0.52	1.72	0.57
Temporal	0.00	2.20	0.72	0.66	0.18	0.73	0.34	0.26	0.86	0.38
Cognitive	0.00	3.08	0.00	0.50	0.18	0.73	0.34	0.26	0.86	0.38
Linguistic	0.00	2.20	0.00	0.17	0.09	0.36	0.17	0.13	0.43	0.19
Total per 1,000 words	24.90	25.07	45.55	34.09	11.47	41.48	21.71	17.59	**54.22**	17.78

Table A.4: Normalized relation distribution (relations per 1,000 words)

disciplines, with Computer Science showing the highest raw count of relations overall.

When normalized for text length, shown in Table A.4, Health/Medicine exhibited the highest relation density (54.22 relations per 1,000 words), followed by Philosophy (45.55) and Political Science (41.48).

A.2 Extraction methodology

The following illustrates the key components of the prompting strategy used for concept and relation extraction. These templates guided the large language model in identifying educationally relevant concepts and relationships.

A.2.1 Concept extraction prompt

```
prompt = f""" 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. The extracted concepts will be used for relation extraction and creating layered visualizations that support flexible, non-linear educational comprehension.

**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
- Critical processes and mechanisms central to understanding

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

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

```

```
**Extraction Guidelines:**
- Tag each extracted concept with its appropriate layer (priority, secondary, tertiary)
- Ensure comprehensive coverage across all layers
- Include concepts that answer: "What" (definitions and principles), "How" (processes and methods), "Why" (reasoning and implications) and"When" (temporal and contextual factors)
- ONLY exclude purely anecdotal details unless they are crucial for defining a concept

**Output Format:**
[
    {{
    "entity": "main_form",
    "context": "The exact sentence where this concept appeared",
    "evidence: "Why this concept is essential for understanding the topic",
    "layer": "priority/secondary/tertiary" # Must be exactly one of these values
    }}
]

Section text:
{full_section_text}
"""

try:
    response = self._cached_api_call(prompt)
    entities = json.loads(OptimizedEntityExtractor.clean_markdown_json(response))
    # Cache results
    self.memory_cache[full_section_text] = entities
    self.cache_manager.cache_entities(full_section_text, entities)

    return entities
```

A.2.2 Concept linking prompt

```
prompt = f"""
Compare these two lists of concepts and identify which ones represent EXACTLY the same abstract idea or unit of knowledge.
If a concept in List 2 matches one in List 1, it should be treated as a variant of that concept.

Guidelines for matching:
1. Match concepts that:
    - Refer to exactly the same concept
    - Are synonyms or alternative expressions
    - Mean the same thing in different contexts

2. Do NOT match concepts that:
    - Are merely related or connected (e.g., "tardigrade anatomy" is not equal to "tardigrade")
    - Have a hierarchical relationship
    - Represent different aspects of the same topic

```

```
Return a simple dictionary mapping concepts from List 2 to their matches in List 1.
If no match exists, don't include that concept.

Example output format:
{json.dumps(sample_output, indent=2)}

List 1:
{json.dumps(normalized_list1, indent=2)}

List 2:
{json.dumps(normalized_list2, indent=2)}
"""

try:
    response = self._cached_api_call(prompt)
    matches = json.loads(self.clean_markdown_json(response))
    original_case_matches = {}
    for new_entity in list2:
        if new_entity["entity"].lower() in matches:
            # Find original case in list1
            for orig_entity in list1:
                if orig_entity["entity"].lower() == matches[new_entity["entity"].lower()]:
                    original_case_matches[new_entity["entity"]] = orig_entity["entity"]
                    break
    return original_case_matches
except Exception as e:
    logger.error(f"Error in concept linking: {e}")
    return {}
```

A.2.3 Local relation extraction prompt

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

**Context:**
The extracted relations should represent meaningful connections that contribute to understanding the main ideas in the text.

**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.

Available Concepts:
{json.dumps([c["id"] for c in concepts], indent=2)}

**Output Format:**
{{
    "relations": [
        {{
            "source": "source concept",
            "relation_type": "type of relationship",
            "target": "target concept",
            "evidence": "text evidence for this relationship"
        }}
    ]
}}

Section Text:
{text}
"""

try:
    response = self._cached_api_call(prompt)
    relations_data = json.loads(self.clean_markdown_json(response))

    relations = []
    for rel in relations_data["relations"]:
        relation = Relation(
            source=rel["source"],
            relation_type=rel["relation_type"],
            target=rel["target"],
            evidence=rel["evidence"],
            section_index=section_info["section_index"],
            section_name=section_info["section_name"]
        )
        relations.append(relation)

    return relations
```

```
except Exception as e:
    logger.error(f"Error extracting local relations: {e}")
    return []
```

A.2.4 Global relation extraction prompt

```
prompt = f"""
Extract global relationships using all processed concepts. The focus is on identifying high-level connections that span across sections or paragraphs, providing a comprehensive understanding of how concepts interrelate on a broader scale.

**Context:**
The extracted global relationships should illustrate overarching connections that tie together multiple sections, enhancing the reader's comprehension of the text as a whole.

**Guidelines:**
- Identify relationships that are significant at a higher level, beyond individual sections or paragraphs.
- Include relationships that show how concepts influence each other across different contexts or sections.
- Ensure each identified relationship is supported by reasoning or textual evidence, highlighting the connection's relevance to the overall content.

Available Concepts:
{json.dumps([c["id"] for c in master_concepts], indent=2)}

```

```
Return in JSON format:
{{
    "relations": [
        {{
            "source": "source concept",
            "relation_type": "type of relationship",
            "target": "target concept",
            "evidence": "reasoning for this relationship"
        }}
    ]
}}
"""

try:
    response = self._cached_api_call(prompt)
    relations_data = json.loads(self.clean_markdown_json(response))

    relations = []
    for rel in relations_data["relations"]:
        relation = Relation(
            source=rel["source"],
            relation_type=rel["relation_type"],
            target=rel["target"],
            evidence=rel["evidence"],
            section_index=-1,  # Indicates global relation
            section_name="global"
        )
        relations.append(relation)

    return relations
except Exception as e:
    logger.error(f"Error extracting global relations: {e}")
    return []
```

Full implementation details, including complete prompt templates and processing algorithms, are available in the project repository at https://github.com/mollyhan19/simplified-cognitext.

Bibliography

[1] D. Anastasiou, C. N. Wirngo, and P. Bagos. The effectiveness of concept maps on students achievement in science: A meta-analysis. *Educational Psychology Review*, 36 (39):1–18, 2024. doi: 10.1007/s10648-024-09877-y. URL `https://doi.org/10.1007/s10648-024-09877-y`.

[2] Dimitrios Anastasiou, Chiawa N. Wirngo, and Pantelis Bagos. The effectiveness of concept maps on students achievement in science: A meta-analysis. *Educational Psychology Review*, 36:39, 2024. doi: 10.1007/s10648-024-09877-y. URL `https://doi.org/10.1007/s10648-024-09877-y`.

[3] Nicholas Asher and Alex Lascarides. Lexical disambiguation in a discourse context. *Journal of Semantics*, 12:69–108, 1995. doi: https://doi.org/10.1093/jos/12.1.69.

[4] David P. Ausubel. *Educational Psychology: A Cognitive View*. Holt, Rinehart & Wilson, 1968.

[5] Ruhil Amal Azmuddin Azmuddin, Nor Fariza Mohd Nor, and Afendi Hamat. Facilitating online reading comprehension in enhanced learning environment using digital annotation tools. *IAFOR Journal of Education*, 8(2):7–27, 2020.

[6] A. Bahari, S. Wu, and P. Ayres. Improving computer-assisted language learning through the lens of cognitive load. *Educational Psychology Review*, 35:53, May 2023. doi: 10.1007/s10648-023-09764-y. URL `https://doi.org/10.1007/s10648-023-09764-y`.

[7] Lawrence W Barsalou. Simulation, situated conceptualization, and prediction. *Philosophical transactions of The Royal Society B: biological sciences*, 364(1521):1281–1289, 2009.

[8] Gal Ben-Yehudah and Adi Brann. 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, 2019. ISSN 0891-4222. doi: 10.1016/j.ridd.2019.04.001. URL `https://www.sciencedirect.com/science/article/pii/S0891422219300605`.

[9] Mohd Nur Hifzhan bin Noordan and Melor Md Yunus. Using digital comprehension to improve reading comprehension skills among young learners. *International Journal of Academic Research in Progressive Education and Development*, 11(2), 2022.

[10] H. Chau, I. Labutov, K. Thaker, et al. Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*, 31:820–846, 2021. doi: 10.1007/s40593-020-00207-1. URL `https://doi.org/10.1007/s40593-020-00207-1`.

[11] O. Chen, F. Paas, and J. Sweller. Cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 35(63):1–18, 2023. doi: 10.1007/s10648-023-09782-w.

[12] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 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, April 25–29 2022. ACM. doi: 10.1145/3485447.3511998. URL `https://doi.org/10.1145/3485447.3511998`.

[13] Zhenbin Chen, Zhixin Li, Yufei Zeng, Canlong Zhang, and Huifang Ma. Gap: A novel generative context-aware prompt-tuning method for relation extraction. *Expert Systems*

with Applications, 248:123478, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.123478. URL https://www.sciencedirect.com/science/article/pii/S0957417424003439.

[14] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs, 2019. URL https://arxiv.org/abs/1909.00228.

[15] Gwo-Jen Hwang Chun-Chun Chang and Yun-Fang Tu. Roles, applications, and trends of concept map-supported learning: a systematic review and bibliometric analysis of publications from 1992 to 2020 in selected educational technology journals. *Interactive Learning Environments*, 31(9):5995–6016, 2023. doi: 10.1080/10494820.2022.2027457. URL https://doi.org/10.1080/10494820.2022.2027457.

[16] Nino B Cocchiarella. Conceptual realism as a formal ontology. In *Formal ontology*, pages 27–60. Springer, 1996.

[17] Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

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

[19] P. Gao, J. Li, and S. Liu. An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. *Mobile Networks and Applications*, 26: 2123–2126, October 2021. doi: 10.1007/s11036-021-01777-7.

[20] Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. Explainable prediction of text complexity: The missing preliminaries for text simplification. In

Proceedings of [Conference Name or Journal], 2024. URL https://arxiv.org/pdf/2007.15823.

[21] David Hay, Ian Kinchin, and Sarah Lygo-Baker. Making learning visible: The role of concept mapping in higher education. *Studies in Higher Education*, 33(3):295–311, 2008. doi: 10.1080/03075070802049251.

[22] 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. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, 01 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad259. URL https://doi.org/10.1093/jamia/ocad259.

[23] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 07 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00324. URL https://doi.org/10.1162/tacl_a_00324.

[24] M.M. Ko-Januchta, K.J. Schnborn, C. Roehrig, et al. connecting concepts helps put main ideas together: cognitive load and usability in learning biology with an ai-enriched textbook. *International Journal of Educational Technology in Higher Education*, 19: 11, March 2022. doi: 10.1186/s41239-021-00317-3. URL https://doi.org/10.1186/s41239-021-00317-3.

[25] Satoshi Kume and Kouji Kozaki. Extracting domain-specific concepts from large-scale linked open data. *CoRR*, abs/2112.03102, 2021. URL https://arxiv.org/abs/2112.03102.

[26] Anne-Laure Le Cunff, Vincent Giampietro, and Eleanor Dommett. Neurodiversity positively predicts perceived extraneous load in online learning: A quantitative research

study. *Education Sciences*, 14(5), 2024. ISSN 2227-7102. doi: 10.3390/educsci14050516. URL https://www.mdpi.com/2227-7102/14/5/516.

[27] Anne-Laure Le Cunff, Vincent Giampietro, and Eleanor Dommett. Neurodiversity and cognitive load in online learning: A focus group study. *Plos one*, 19(4):e0301932, 2024.

[28] Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. Pushing on text readability assessment: A transformer meets handcrafted linguistic features, 2024. URL https://arxiv.org/abs/2109.12258.

[29] Sina Lenski, Stefanie Elsner, and Jrg Groschedl. Comparing construction and study of concept maps an intervention study on learning outcome, self-evaluation and enjoyment through training and learning. *Frontiers in Education*, 7, 2022. ISSN 2504-284X. doi: 10.3389/feduc.2022.892312. URL https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2022.892312.

[30] Alvin Ping Leong. Marked themes in academic writing: a comparative look at the sciences and humanities. *Text & Talk*, 0, 2024. URL https://api.semanticscholar.org/CorpusID:267192615.

[31] Yifan Li and Lingling Yan. Which reading comprehension is better? a meta-analysis of the effect of paper versus digital reading in recent 20 years. *Telematics and Informatics Reports*, 14:100142, 2024. ISSN 2772-5030. doi: https://doi.org/10.1016/j.teler.2024.100142. URL https://www.sciencedirect.com/science/article/pii/S2772503024000288.

[32] Hsin-Yi Liang, Tien-Yu Hsu, and Gwo-Jen Hwang. Promoting children's inquiry performances in alternate reality games: A mobile concept mapping-based questioning approach. *British Journal of Educational Technology*, 52(5):2000–2019, September 2021. doi: 10.1111/bjet.13095. URL https://doi.org/10.1111/bjet.13095.

[33] R. Liu, T. Cheng, and L. Zhou. A comparative survey of online reading and paper reading behavior. *E-Education Research*, 05:28–31, 2004. doi: 10.13811/j.cnki.eer.2004.05.006.

[34] María J López-Huertas. Domain analysis for interdisciplinary knowledge domains. *KO KNOWLEDGE ORGANIZATION*, 42(8):570–580, 2015.

[35] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv*, abs/1808.09602, 2018. URL `https://arxiv.org/abs/1808.09602`.

[36] Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. CNER: Concept and named entity recognition. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.461. URL `https://aclanthology.org/2024.naacl-long.461/`.

[37] Richard E. Mayer. Cognitive theory of multimedia learning. In Richard E. Mayer, editor, *The Cambridge Handbook of Multimedia Learning*, pages 43–71. Cambridge University Press, 2nd edition, 2014. doi: 10.1017/CBO9781139547369.005.

[38] Mohammadreza Molavi, MohammadReza Tavakoli, and G'abor Kismih'ok. Extracting topics from open educational resources. In *European Conference on Technology Enhanced Learning*, 2020. URL `https://api.semanticscholar.org/CorpusID:219956251`.

[39] Adeladlew Kassie Netere, Anna-Marie Babey, Roisin Kelly-Laubscher, Thomas A. Angelo, and Paul J. White. Mapping design stages and methodologies for developing stem concept inventories: a scoping review. *Frontiers in Education*, 2024. URL `https://api.semanticscholar.org/CorpusID:273881518`.

[40] Joseph Novak and Alberto Caas. Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting Education*, 3, 01

2007. URL https://www.informationtamers.com/PDF/Theoretical_origins_of_concept_maps,_how_to_construct_them,_and_uses_in_education.pdf.

[41] Kaitlyn M.A. Parks, Christine N. Moreau, Kara E. Hannah, Leah Brainin, and Marc F. Joanisse. The task matters: A scoping review on reading comprehension abilities in adhd. *Journal of Attention Disorders*, 26(10):1304–1324, 2022. doi: 10.1177/10870547211068047. URL https://doi.org/10.1177/10870547211068047. PMID: 34961391.

[42] Han Qin, Yuanhe Tian, and Yan Song. Enhancing relation extraction via adversarial multi-task learning. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6190–6199, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.666/.

[43] Frank Safayeni, Natalia Derbentseva, and Alberto Caas. Concept maps: A theoretical note on concepts and the need for cyclic concept maps. *Journal of Research in Science Teaching*, 42:741 – 766, 09 2005. doi: 10.1002/tea.20074.

[44] N.L. Schroeder, J.C. Nesbit, C.J. Anguiano, et al. Studying and constructing concept maps: a meta-analysis. *Educational Psychology Review*, 30:431–455, 2018. doi: 10.1007/s10648-017-9403-9. URL https://doi.org/10.1007/s10648-017-9403-9.

[45] T. Shealy, J. S. Gero, and P. Ignacio. How the use of concept maps changes students' minds and brains. In *ASEE Annual Conference and Exposition, Conference Proceedings*, 2022. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138317548&partnerID=40&md5=b58c70083e60312be6bfbf2c0bedb301.

[46] A. Skulmowski and K.M. Xu. Understanding cognitive load in digital and online learning:

a new perspective on extraneous cognitive load. *Educational Psychology Review*, 34: 171–196, March 2022. doi: 10.1007/s10648-021-09624-7.

[47] Junyoung Son, Jinsung Kim, Jungwoo Lim, and Heuiseok Lim. Grasp: Guiding model with relational semantics using prompt for dialogue relation extraction, 2022. URL `https://arxiv.org/abs/2208.12494`.

[48] Robyn Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `https://aclanthology.org/L12-1639/`.

[49] L. Sperotto. 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, 2016. doi: 10.1080/1034912X.2016.1153256. URL `https://doi.org/10.1080/1034912X.2016.1153256`.

[50] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. doi: 10.1016/0364-0213(88)90023-7. URL `https://doi.org/10.1016/0364-0213(88)90023-7`.

[51] Nairn Ta and Abu Bakar Razali. Concept mapping for improving reading comprehension in second language education: A systematic review. *International Journal of Learning, Teaching and Educational Research*, 2023. URL `https://api.semanticscholar.org/CorpusID:261453634`.

[52] Omar Taky-eddine and Redouane Madaoui. Cognitive overload in the hypertext reading environment. *International Journal of English Language Studies*, 6(2):94, May 2024. ISSN

2707-7578. doi: 10.32996/ijels.2024.6.2.13. URL https://www.al-kindipublisher.com/index.php/ijels.

[53] Aydin Durgunoglu Turkan Ocal and Lauren Twite. Reading from screen vs reading from paper: Does it really matter? *Journal of College Reading and Learning*, 52(2): 130–148, 2022. doi: 10.1080/10790195.2022.2028593. URL https://doi.org/10.1080/10790195.2022.2028593.

[54] Stephan van Gasselt and Andrea Nass. A semantic view on planetary mapping - investigating limitations and knowledge modeling through contextualization and composition. *Remote. Sens.*, 15:1616, 2023. URL https://api.semanticscholar.org/CorpusID:257610853.

[55] Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 23612364, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358119. URL https://doi.org/10.1145/3357384.3358119.

[56] KH. Yang, HC. Chu, and GJ. Hwang. A progressive concept map-based digital gaming approach for mathematics courses. *Educational Technology Research and Development*, 2025. doi: 10.1007/s11423-025-10461-6. URL https://doi.org/10.1007/s11423-025-10461-6. Accepted: 31 January 2025, Published: 14 February 2025.

[57] M. Yeari, E. Vakil, L. Schifer, and R. Schiff. The origin of the centrality deficit in individuals with attention-deficit/hyperactivity disorder. *Journal of Clinical and Experimental Neuropsychology*, 41(1):69–86, 2018. doi: 10.1080/13803395.2018.1501000. URL https://doi.org/10.1080/13803395.2018.1501000.

[58] Z Yuan and X Bai. Watching computer screens is different from reading books. *Research in Educational Development*, 36(20):15–20, 2016.

[59] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 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, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581260. URL `https://doi.org/10.1145/3544548.3581260`.

[60] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11), July 2024. ISSN 0360-0300. doi: 10.1145/3674501. URL `https://doi.org/10.1145/3674501`.

[61] Michal Zivan, Sasson Vaknin, Nimrod Peleg, Rakefet Ackerman, and Tzipi Horowitz-Kraus. 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, 2023.