Creating Local World Models using LLMs

Mark David Longar
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Novak
Jožef Stefan Institute
Ljubljana, Slovenia

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

A key limitation of state-of-the-art large language models is their lack of a consistent world model, which hinders their ability to perform unseen multi-hop reasoning tasks. This paper addresses this by extracting local world models from text into a systematic first-order logic framework, enabling structured reasoning. Focusing on the educational domain, we present a multi-step approach using Prolog to represent and reason with these models. Our method involves segmenting educational texts, generating Prolog definitions, and merging them into a comprehensive knowledge graph. We successfully extracted several small models and manually verified their accuracy, demonstrating the potential of this approach. While promising, our results are currently limited to small-scale models.

Keywords

Large language models, local world models, knowledge representation, educational technology, structured reasoning, knowledge graphs

1 Introduction

In recent years, Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), offering unprecedented capabilities in understanding, reasoning over, and generating human-like text. Despite their impressive performance across various language tasks, a significant limitation persists – the absence of a consistent and coherent world model within these systems [8]. This limitation hampers their ability to perform advanced reasoning tasks that require not only textual understanding but also logical consistency and structured knowledge representation.

While current LLMs are powerful, they are inherently constrained by their reliance on statistical correlations within vast datasets, often resulting in shallow and contextually inconsistent reasoning. To address this limitation, we propose an approach for extracting local world models, i.e., small, context-specific representations of knowledge that capture the relationships and rules governing a particular domain or scenario. The approach is multi-step. First, the input text is segmented into manageable parts. Each segment is analyzed to extract key concepts and their interrelationships, which are then represented as Prolog definitions. Then, the definitions are merged into a comprehensive knowledge graph that reflects the structure and content of the input text.

We focus specifically on the educational domain, where the ability to generate and utilize local world models could significantly enhance the effectiveness of AI-driven educational tools. We applied the approach to two textbooks to show its capabilities of creating logically coherent and pedagogically sound knowledge graphs. However, by modifying some of the components, the approach can also be applied to other domains, such as industry, finance, and law.

The remainder of the paper is as follows: Section 2 presents the related work on LLMs and creating world models. Next, the proposed approach is described in Section 3. The experiment setting is presented in Section 4, followed by the experiment results in Section 5. We discuss the results in Section 6 and conclude the paper in Section 7.

2 Related Work

The recent surge in large language models, such as GPT-3 [3] and GPT-4 [1], has significantly advanced natural language processing, showing emergent reasoning abilities across various tasks. However, despite their impressive performance, LLMs are often criticized for lacking factual consistency, interpretability, and logical coherence, especially in complex, multi-hop reasoning tasks [8]. To address these shortcomings, efforts have been made to integrate LLMs with structured knowledge frameworks, like knowledge graphs (KGs) and ontologies, to enhance reasoning and knowledge flow between structured data and language models [9].

In the field of ontology and KG development, early initiatives like Cyc [6] laid the groundwork for large-scale structured knowledge representation. More recent efforts [8, 5] have explored using LLMs to assist in ontology generation and KG construction. While LLMs can automate parts of the ontology development process, they struggle with ensuring logical consistency and managing complex domain-specific knowledge [5, 2]. Complementary approaches, like using LLMs for ontology learning [2] and structured knowledge extraction [10], highlight the need for human validation and formal methods to ensure accuracy.

Our work builds on these insights by focusing on using LLMs to extract structured local world models in the form of Prolog-based representations. This approach addresses the limitations of LLMs in handling complex reasoning and provides a more robust, logically consistent framework for educational applications.

3 Methodology

This section introduces the approach for creating local world models by generating and utilizing structured data in Prolog. The methodology is designed to systematically identify and map the concepts and their interrelationships within a given educational document, such as a textbook, facilitating the generation of a knowledge graph.

3.1 Document segmentation

To manage the document's complexity and ensure accurate concept extraction, the source material was divided into several shorter parts, each up to 10 pages long. This segmentation was crucial in allowing us to focus on smaller, more manageable sections of the content, enabling a thorough analysis and avoiding problems that come with long-context LLM outputs. The length

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

of each part was determined based on the natural divisions within the text, such as chapters or major sections, to maintain the coherence of concepts within each segment.

3.2 Generating Prolog definitions

For each segmented part, we created a prompt to generate Prolog definitions of the concepts and their relationships. The prompt was carefully crafted to guide the extraction of educational content in a structured format. It consisted of three main components: the context, the predicates and the structured output.

Context. A description of the educational context and a brief narrative to position the content within a learning scenario. This helped to align the LLM-extracted concepts and relationships with our downstream tasks. The following is an example of the prompt used:

> *You are a teacher and an expert in natural language processing (NLP). You wrote a chapter in an NLP textbook and would like to convert the content of the chapter into a classroom lesson. You would like to step into the shoes of a student in order to understand their learning process of this material. You need to understand which concepts are being taught and their relationships.*

Predicates. List of predicates and their descriptions, which were essential for identifying concepts (`isConcept(A)`), prerequisites (`isPrerequisiteOf(A, B)`), and sections (`isSection(S)`). These predicates were used to simulate the learning process, where concepts are linked to sections. A concept may have prerequisite concepts or sections that must be understood before a student can advance to learning the concept.

Structured output. Clear instructions to output the extracted predicates in the form of a Prolog program. The LLM responding in a structured format a crucial part of our approach, as it has been shown that structured responses can improve LLM reasoning and generation quality [12].

In summary, this prompt allowed us to extract detailed summaries of the concepts taught and their relationships, which were then represented in Prolog. Each segment was processed independently to generate a corresponding Prolog program.

3.3 Merging Prolog definitions

After generating the Prolog definitions for each segment, the next step was to merge them into a single cohesive program. To achieve this, we created a prompt, which instructed the system to combine the three disjoint parts into one integrated Prolog program. This process involved:

- **Incorporation of All Concepts.** Ensuring that all concepts and relationships from the segmented parts were included in the final program.
- **Integration of Sections.** Merging related sections and establishing connections between different parts of the knowledge graph.
- **Refinement.** Adjusting the relationships and concepts to ensure logical consistency across the knowledge graph.

The final output was a comprehensive Prolog program that accurately reflected the educational content and its structure.

3.4 Use of the knowledge graph

The generated knowledge graph, represented by the Prolog program, was then used to recommend the next steps in the learning process. Using the structured output, we created a detailed concept map that helped identify key learning paths and prerequisites. Prolog was chosen for this task because it can handle structured data, is widely used (increasing the likelihood that LLMs have encountered it during training), and can be executed and analyzed immediately.

4 Experiment Setting

This section outlines the experiment setting for evaluating our approach to extracting local world models from educational texts and generating structured Prolog representations. We describe the data sources, the large language model used, and the evaluation framework.

4.1 Data sources

We evaluated our approach on two widely used textbooks in deep learning and natural language processing. These texts were chosen because they are relevant to both structured reasoning tasks and the representation of complex, multi-step concepts. The following chapters were selected for analysis:

Deep Learning Preliminaries from the book *Dive into Deep Learning* [11]. This chapter provides foundational knowledge of deep learning, covering key concepts such as linear algebra, calculus, and probability, which are essential for understanding the field. The textbook's teaching approach is highly hands-on, with a significant portion devoted to code. It is open-sourced, and we used the Markdown files provided on their GitHub page[1].

Chapter 2: Regular Expressions, Tokenization, and Edit Distance from *Speech and Language Processing* [4]. This chapter introduces basic NLP techniques, focusing on regular expressions and tokenization, which are pivotal in text preprocessing tasks.

4.2 Used large language model

We employed GPT-4o via the ChatGPT interface to extract concepts and their interrelationships. We leveraged the model's multimodal capabilities, allowing it to process text and PDF documents.

4.3 Evaluation framework

We developed an evaluation framework to measure the performance of our approach based on three primary metrics: accuracy, completeness, and consistency. Additionally, we employed manual verification to assess the correctness of the extracted knowledge models.

Metrics for Model Evaluation. The following metrics were used to evaluate the effectiveness of our approach:

- *Accuracy.* This metric measures how accurately the approach extracted the concepts and their relationships from the text. We evaluated the correctness of each Prolog definition against the source material.
- *Completeness.* This evaluates whether the system captured all the key concepts from the educational material. We ensured

[1] https://github.com/d2l-ai/d2l-en

that no significant concepts or relationships were omitted during extraction.

- *Consistency.* The metric assesses the extent to which the extracted models maintained logical coherence across different segments of the text. This was crucial in determining whether the segmented Prolog definitions could be merged into a cohesive KG.

Ground truth verification. To validate the results, we employed deep learning and NLP experts who manually reviewed the extracted knowledge graphs and compared them with the source texts. They ensured that the extracted concepts were accurate, complete, and logically consistent. Any discrepancies were noted and used to refine the extraction process, providing a feedback loop for improving future iterations of the model.

5 Results

In this section, we review the knowledge graphs of the two tested texts generated by our model.

5.1 *Dive into Deep Learning*

The selected chapter covered six sub-chapters in the following order: Data Manipulation, Data Preprocessing, Linear Algebra, Calculus, Automatic Differentiation, and Probability and Statistics. The results are represented by the graph in Figure 1.

The system accurately identified three major independent branches of the chapter – Linear Algebra, Calculus, and Probability and Statistics – which reflects the structure of the source material. The extracted knowledge graph also logically restructured the content in ways that differed from the original organization but made sense pedagogically. This restructuring highlights the logical flow of how data handling techniques naturally feed into more abstract mathematical concepts despite differing from the original structure.

However, some omissions and reassignments were noted, particularly within the Linear Algebra section. Concepts such as vectors and matrices were omitted, likely due to the high-level nature of the extraction process. Additionally, matrix multiplication, though identified, was separated from Linear Algebra basics and Tensor operations. This disjunction represents a slight deviation from the expected conceptual hierarchy.

Similarly, in the Calculus section, the extracted model restructured the sequence of topics. This restructuring captured the relationship between fundamental calculus concepts and their practical applications in machine learning. Furthermore, the system included concepts like Gradient Descent and Backpropagation which were only briefly mentioned in the source material.

5.2 *Speech and Language Processing*

The Regular Expressions section, seen in Figure 2, was extracted accurately, capturing the core concepts effectively. However, a noticeable limitation was the loss of the original sequencing of the concepts presented in the textbook. While the key ideas were identified, the pedagogical flow, which is essential for gradual learning, was somewhat disrupted in the extraction process.

For the other sections, including Tokenization and Edit Distance, the model extracted only the most prominent concepts, omitting many important details. As a result, these sections are less comprehensive than they need to be for in-depth understanding. Despite this, the overall connections between sections in the knowledge graph were logically structured, showing that the system was still able to create a coherent representation of the material at a high level.

It is important to note that this textbook is significantly more information-dense and longer compared to the *Dive into Deep Learning* book. This added complexity exposed some limitations in the current approach, mainly when dealing with texts that require detailed extraction of concepts and their interrelationships. The model's ability to handle such dense material is limited by its tendency to focus on top-level ideas while losing much of the depth and sequencing provided in the source text.

6 Discussion

Our approach to extracting local world models from educational texts demonstrated strong performance in generating logically coherent knowledge graphs from high-level concepts, but certain limitations were identified. The synthetic data generation effectively captured core concepts from both textbooks, particularly in structuring major branches such as Linear Algebra, Calculus, and Probability from *Dive into Deep Learning*. However, some restructured sections, while logical, differed significantly from the source material's flow.

In the *Speech and Language Processing* textbook, Regular Expressions were extracted with sufficient accuracy. Other sections, such as Tokenization and Edit Distance, suffered from detail omissions, where only top-level concepts were extracted. This issue was more prominent due to the higher information density of the NLP textbook, exposing limitations in handling detailed, densely packed content.

Regarding the evaluation framework, the model generally performed well on metrics like accuracy and consistency but struggled with completeness in more detailed sections. The model's tendency to restructure content logically, though sometimes deviating from the original, suggests that while it captures core relationships, further refinements are needed to preserve pedagogical flow and details.

6.1 Potential improvements

To address the limitations, improving the prompt engineering could lead to more detailed extractions while maintaining the structure of the source material. Additionally, enhancing the model's ability to handle complex, dense information would mitigate the loss of key concepts. Future iterations may benefit from automated post-processing checks to ensure logical consistency and reduce manual interventions. Overall, while the approach shows promise, refining it to handle finer details and complex sequences more effectively will be essential for broader applications.

7 Conclusion and Future work

In this paper, we proposed a novel approach to extracting local world models from educational texts by generating structured Prolog representations. Our methodology demonstrated the ability to capture core concepts and their interrelationships in a logical and coherent manner, especially in the *Dive into Deep Learning* textbook. However, the results from the more information-dense *Speech and Language Processing* text revealed limitations, particularly in handling detailed content, large knowledge graphs, as well as preserving pedagogical flow.

The use of Prolog proved effective in organizing educational material, allowing for structured reasoning and enabling applications in AI-driven educational tools. Despite these successes,

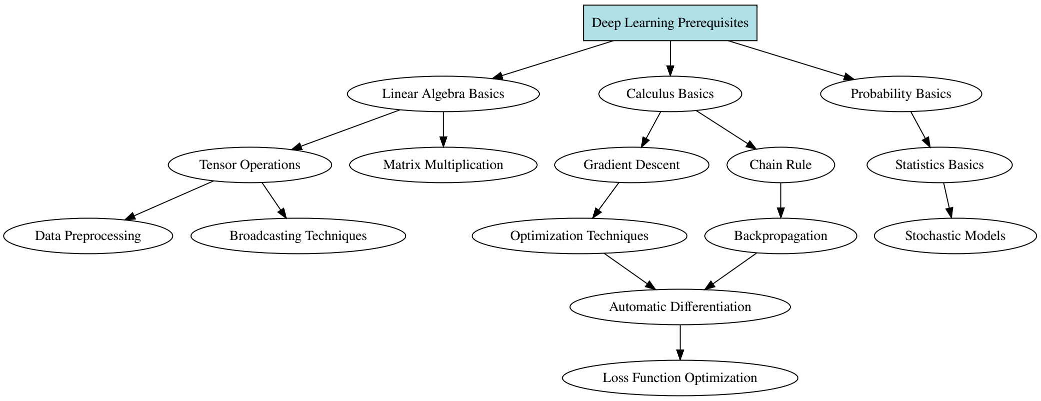

Figure 1: Knowledge graph of the *Preliminaries* section from *Dive into Deep Learning*.

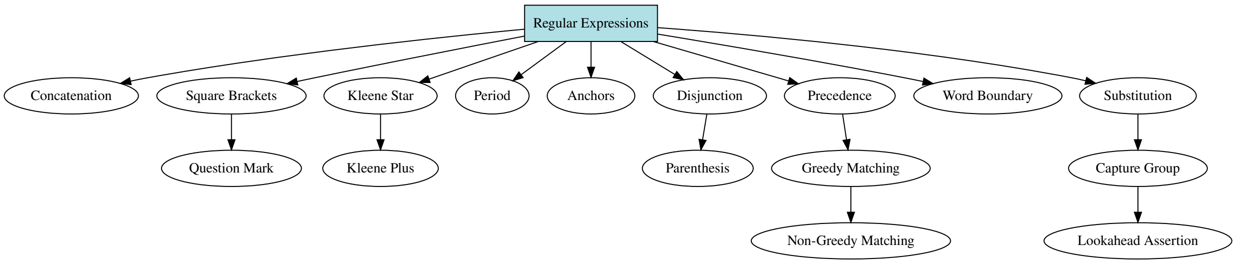

Figure 2: Knowledge graph of the *Regular Expressions* section from *Speech and Language Processing*.

certain challenges remain, such as the omission of detailed concepts and the system's occasional tendency to deviate from the original sequence of topics.

Future work will address these limitations by improving the prompt engineering and enhancing the system's ability to handle complex, information-dense material. Additionally, we plan to explore automating the segmentation process and scaling up the model to generate larger, more intricate knowledge graphs. Other potential directions include integrating retrieval-augmented generation [7] to enrich knowledge extraction and comparing generated world models across different texts to evaluate their pedagogical alignment. Self-evaluation and correction mechanisms could also be introduced to improve accuracy and completeness.

Acknowledgments

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 project Humane AI Net (Grant No. 952026).

References

[1] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. "LLMs4OL: Large language models for ontology learning". In: *International Semantic Web Conference*. Springer. 2023, pp. 408–427.

[3] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (visited on 08/27/2024).

[4] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released August 20, 2024. 2024. URL: https://web.stanford.edu/~jurafsky/slp3/.

[5] Vamsi Krishna Kommineni, Birgitta König-Ries, and Sheeba Samuel. "From human experts to machines: An LLM supported approach to ontology and knowledge graph construction". In: *arXiv preprint arXiv:2403.08345* (2024).

[6] Douglas B Lenat. "CYC: A large-scale investment in knowledge infrastructure". In: *Communications of the ACM* 38.11 (1995), pp. 33–38.

[7] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[8] Fabian Neuhaus. "Ontologies in the era of large language models–a perspective". In: *Applied ontology* 18.4 (2023), pp. 399–407.

[9] Shirui Pan et al. "Unifying large language models and knowledge graphs: A roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* (2024).

[10] Mohammad Javad Saeedizade and Eva Blomqvist. "Navigating Ontology Development with Large Language Models". In: *European Semantic Web Conference*. Springer. 2024, pp. 143–161.

[11] Aston Zhang et al. *Dive into Deep Learning*. https://D2L.ai. Cambridge University Press, 2023.

[12] Pei Zhou et al. "How FaR Are Large Language Models From Agents with Theory-of-Mind?" In: *arXiv preprint arXiv:2310.03051* (2023).