Using Large Language Models to Convert Documents to Knowledge Graphs to Check for Completeness and Consistency

by Michael Wacey

B.S. in Computer Science, June 1985, Washington University in St. Louis
M.S. in Computer Science, December 1990, Drexel University

A Dissertation submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Doctor of Philosophy

March 8, 2026

Dissertation directed by

Professor M. Elbasheer
Professor of Department of Computer Science

The School of Engineering and Applied Science of The George Washington University certifies that Michael Wacey has passed the Final Examination for the degree of Doctor of Philosophy as of December 18, 2025. This is the final and approved form of the dissertation.

Using Large Language Models to Convert Documents to Knowledge Graphs to Check for Completeness and Consistency

Michael Wacey

Dissertation Research Committee:

Professor M. Elbasheer, Professor of Department of Computer Science, Dissertation Director

Full Name, Title, Dissertation Director/Dissertation Co-Director/Committee Member

Full Name, Title, Dissertation Director/Dissertation Co-Director/Committee Member

Full Name, Title, Dissertation Director/Dissertation Co-Director/Committee Member

Name of External Examiner, Professorial Title, Name of External University (or Name, Job Title, Name of External Company), Committee Member

Dedication

Include a fancy quote or dedication.

Acknowledgments

Here you can acknowledge all of those people who have helped you to reach this point. It's rare that any work is done in a vacuum and your research is no exception. Feel free to be grateful for all those who've aided you along your way.

Abstract

Using Large Language Models to Convert Documents to Knowledge Graphs to Check for Completeness and Consistency

This is the abstract. It contains some random text from the `lipsum` package. You may safely remove the `lipsum` package once you write your thesis.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Table of Contents

List of Figures

List of Tables

Nomenclature

Preface

This is the preface. It's another front matter page that offers additional detail into your work. Typically, only one (preface OR prologue OR foreword) is used. You can remove the other sections by deleting them inside `tex/frontmatter.tex` or using the appropriate show or hide commands.

Chapter 1: Introduction

1.1 Background

Ensuring document quality involves verifying completeness, consistency, and correctness. While evaluating correctness often necessitates external knowledge and understanding the document's intent, completeness and consistency can be assessed within the document itself. This research focuses on developing automated methods using Large Language Models (LLMs) to address the latter two aspects. The specific focus is on converting a large document into a knowledge graph that can be used in future research to check the consistency and completeness of a document.

All LLMs are derived from the Attention Model and have a window of information that they can look at. As documents grow larger, they exceed this window and therefore cannot be processed completely by an LLM. Checking a document does not actually require paying attention to the entire document. Instead, the LLM can be used to identify entities of interest throughout the document. Then focus can be applied to the entities that are of a similar type. This combines the best of both attention and focus.

For example, a set of laws may define the various boundaries for plots of land early in the document. This could include minimum frontage, setback from neighbors, and many other items. Each set of boundaries would be for a specific zoning area. At the end of the document the zoning areas are defined. A simple inconsistency would be defining plot boundaries for a zoning area that is never defined. There are many other such interactions between sections of municipal laws that need to be kept synchronized.

1.2 Research Motivation

Despite extensive research on analyzing small documents or specific document sections, there is a significant gap in addressing the challenges of complete, large-scale document analysis. The need for automated consistency and completeness checks is critical in various industries. Currently, these tasks are often performed manually, requiring substantial time and resources while still potentially yielding suboptimal results. This research aims to bridge this gap by developing an effective and efficient automated solution.

Within the scope of this research, local regulations of townships in Pennsylvania go through a time consuming and complicated process to get published. After the governing body enacts a law, it is sent to an organization to compile it into existing laws of the township. This is a manual and intensive process to determine if any of the existing laws are affected by the new law. Even with this, there are many cases of new laws making a set of existing laws incomplete or inconsistent.

1.3 Problem Statement

Municipal laws in Pennsylvania Townships, authored by multiple people over time, develop inconsistencies and are incomplete (D. Curley, Easttown Supervisor, personal communication, September 16, 2024; A. Rau, Esq., Easttown Solicitor, personal communication, September 20, 2024; J. Sanders, personal communication, October 25, 2024), leading to annual revenue losses of hundreds of thousands of dollars. (M. Wacey, Easttown Supervisor, personal communication, September 23, 2024).

1.4 Research Objectives

The primary objective of this research is to develop a tool capable of automatically processing documents of any size into a coherent set of entities in a knowledge graph. This tool will leverage advanced techniques to analyze document content, identify potential entities, and provide access to the knowledge graph.

1.5 Research Questions and Hypotheses

To achieve the research objectives, the following research questions and hypotheses are proposed:

- **RQ1:** Can an LLM be used to convert a large document into a knowledge graph?
- **RQ2:** Can an LLM be used to process multiple knowledge graphs into a typed cluster of knowledge graphs.
- **RQ3:** Can a typed cluster of knowledge graphs be used to check the source document for consistency and completeness?
- **H1:** An LLM can be used to convert a large document into a knowledge graph.
- **H2:** An LLM can be used to process multiple knowledge graphs into a typed cluster of knowledge graphs.
- **H3:** A typed cluster of knowledge graphs can be used to check the source document for consistency and completeness.

1.6 Scope of Research

This research uses Pennsylvania township laws as a case study to train and test the developed tool. These laws, available in the public domain in both PDF and Word formats, are chosen due to their extensive length, multiple authors, and rigorous manual checks for consistency and completeness. Although the focus is on these legal documents, the developed approach is intended to be generalizable to various document types.

For this effort, the development will focus on the creation of Knowledge Graphs that represent the document or documents under review. There will be an analysis to ensure that the knowledge graphs created are suitable to check the document for consistency and completeness. However, that development will be left for future work.

1.7 Research Limitations

This section will be completed later in the research process to accurately reflect any limitations encountered during the development and evaluation of the tool. Potential limitations might include computational constraints, challenges in handling specific document formats or language complexities, and the need for further refinement of neural network models.

1.8 Organization of Praxis

The remainder of this research is organized as follows.

- **Chapter 2 - Literature Review:** This chapter provides a comprehensive review of the relevant literature on:
  - The creation of knowledge graphs from documents by LLMs.

  - The processing of multiple knowledge graphs into a combined knowledge graph by LLMs.
  - The ability of knowledge graphs to be used to represent the original document in ensuring consistency and completeness.
  - The process of creating and maintaining local laws in Pennsylvania.
  - Background on checking documents for consistency and completeness.

- **Chapter 3 - Methodology:** This chapter details the statistical and machine learning methodologies employed in this research, including data pre-processing, model selection, training, and evaluation.
- **Chapter 4 - Results:** This chapter presents and analyzes the results of the data analysis, addressing each research question and hypothesis. It also evaluates the performance of the proposed methodology and tool.
- **Chapter 5 - Discussion and Conclusion:** This chapter concludes the research with a discussion of key findings, contributions to the field, recommendations for practical applications and potential avenues for future research.

Chapter 2: Literature Review

This is the literature review.

2.1 Introduction

This is a section.

Chapter 3: Methodology

3.1 Introduction

Chapter 4: Results

4.1 Introduction

Chapter 5: Conclusions

5.1 Introduction

Bibliography

Appendix A: Appendix

This is an example of an appendix. The only difference is the use of `\appendix` command at the start of this `tex` file. This automatically changes the chapter and section headings.

A.1 A section

The easiest method.

$$x_k = \frac{a_k + b_k}{2} \tag{A.1}$$

A.2 False Position

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

A.3 Starting the Appendices

Actually, using appendices is quite simple. Immediately after the end of the last chapter and before the start of the first appendix, simply enter

the command `\appendix`. This will tell LaTeX to change how it interprets the commands `\chapter`, `\section`, *etc.*

Each appendix is actually a chapter, so once the `\appendix` command has been called, start a new appendix by simply using the `\chapter` command.

Note that the `\appendix` command should be called only once–not before the start of each appendix.

All the fancy referencing and tools still work. You only need to add the appendix command and all will be as it should be.

Appendix B: Another Appendix

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.