Chapter 1: Introduction

1.1 Background and Research Motivation

Ensuring document quality involves verifying completeness, consistency, and correctness (Zowghi & Gervasi, 2003). While evaluating correctness often necessitates access to knowledge external to the document 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.

1.1.1 Background

The increasing complexity and scale of textual documents in various domains present significant challenges in ensuring consistency and completeness. Legal codes, technical documentation, and regulatory frameworks are often drafted collaboratively over extended periods, leading to inconsistencies, redundancies, and gaps in information. Traditional manual review methods, while necessary, are labor intensive and prone to human oversight, making automated solutions an attractive alternative. Advances in natural language processing (NLP) and artificial intelligence (AI) have introduced new methodologies for analyzing and structuring large bodies of text, with promising applications in document validation and knowledge extraction.

At the core of modern NLP advancements are Transformer-based models that rely on the Attention Mechanism to understand and generate text. LLMs,

which build upon this foundation, can process and interpret vast amounts of textual data, though they are constrained by fixed context windows. To address this limitation, structured approaches such as knowledge graphs have emerged, enabling explicit representation of entities and relationships within documents. This research applies these technologies to Pennsylvania township laws, a domain where maintaining consistency is particularly critical. Given the size and complexity of municipal codes, inconsistencies in legal definitions, zoning regulations, and procedural rules can lead to legal disputes and financial losses. By leveraging AI-driven tools, this study aims to develop a framework for systematically analyzing and improving the consistency of legal documents.

Ensuring structural consistency and completeness in documents has been a longstanding challenge in various domains. Previous research has focused on methods to maintain internal coherence within documents (Laban et al., 2021), while other studies have explored domain-specific approaches to consistency checking (Tröls et al., 2022). In academic literature, the term coherence is often used interchangeably with consistency (Shen et al., 2021), reflecting the broader goal of ensuring logical and semantic alignment within textual content.

In 2017, a research team at Google introduced the Transformer model, a neural network architecture based entirely on the Attention Mechanism (Vaswani et al., 2017). Unlike previous sequential models, the Transformer processes all words within a given input simultaneously, allowing it to assess how each word influences others across the text. Using self-attention, this architecture captures long-range dependencies more effectively than previous models. However, despite advances in scaling Transformer-based models, they remain constrained by a limited attention window due to

memory and computational efficiency considerations.

Large Language Models are built upon the Transformer architecture and inherit its fundamental attention-based mechanisms. However, LLMs are constrained by a fixed context window, limiting the amount of text they can analyze at once. As documents grow in length, they often exceed this window, preventing comprehensive processing in a single pass. Despite this limitation, document analysis does not necessarily require attending to an entire document at once. Instead, LLMs can be employed to extract key entities and concepts across different sections, enabling a more focused and structured approach to consistency checking. By identifying entities of interest and analyzing their relationships, LLMs can effectively navigate large documents while maintaining efficiency.

Knowledge graphs provide a structured, human-readable representation of information, serving as an alternative to the implicit encoding of knowledge found in neural networks. A knowledge graph is a directed acyclic graph in which nodes represent entities, and edges define relationships between them. Each node can possess attributes that enrich its descriptive properties. For instance, a node representing a car might include attributes such as color, model, or manufacturer. One useful way to conceptualize knowledge graphs is through the framework of frames, as described by Minsky (Minsky, 1974). Unlike LLMs, which rely on statistical inference, knowledge graphs offer explicit, interpretable relationships that can be leveraged for consistency and completeness verification in structured documents.

Pennsylvania is home to over 1,200 townships of the second class, each responsible for drafting and maintaining its own set of municipal laws. These laws regulate a wide range of local governance areas, including police services, fire departments, zoning, and land development. Over time,

the cumulative nature of legal amendments introduces inconsistencies and gaps, which, if left unaddressed, can lead to legal ambiguities and enforcement challenges. While legal professionals and municipal officials work diligently to identify and resolve these issues, the complexity of these documents—often spanning thousands of pages—makes manual review error-prone and inefficient.

A key source of complexity is the interdependence of different sections within municipal codes. For example, early sections may define zoning regulations, specifying minimum frontage, setbacks, and other boundary constraints for different zoning districts. However, inconsistencies can arise when later sections introduce or reference zoning areas that were never formally defined. Similar discrepancies can emerge across other regulatory provisions, requiring careful synchronization of legal language and definitions. Ensuring consistency across these interconnected legal elements is a critical challenge that demands a more systematic and automated approach to legal document analysis.

1.1.2 Research Motivation

Despite extensive research on the analysis of small documents or specific sections of documents, there is a significant gap in addressing the challenges of comprehensive, 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 be 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 that make a set of existing laws incomplete or inconsistent.

1.2 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).

The complexity of municipal laws in Pennsylvania townships arises from their incremental development over time. Ordinances and regulations are often drafted by different individuals, including elected officials, legal counsel, and administrative staff, each contributing to the evolving legal framework. However, this decentralized process can lead to inconsistencies in language, overlapping provisions, and unintended gaps in regulatory coverage. As laws are amended or new ones are introduced, prior statutes may not be adequately reconciled, further exacerbating these inconsistencies. Without a systematic approach to maintaining legal coherence, townships face challenges in enforcing their laws effectively and equitably.

The consequences of these inconsistencies extend beyond legal ambiguity. Incomplete or conflicting municipal laws can create loopholes that

hinder the township's ability to collect fees, fines, and other sources of revenue. For example, unclear zoning regulations may allow developments to proceed without appropriate permits or impact fees, and ambiguous tax ordinances may lead to disputes that reduce collections. When enforcement mechanisms are weak due to gaps in the legal framework, municipalities struggle to ensure compliance, leading to significant financial losses. These inefficiencies, compounded over time, place additional strain on local budgets, limiting resources for essential public services and infrastructure improvements.

Addressing these issues requires a structured methodology for analyzing, refining, and maintaining municipal laws. Traditional legal review processes, while valuable, are labor-intensive and reactive, often only identifying issues when disputes or financial shortfalls arise. Advances in artificial intelligence, particularly the use of LLMs, offer a potential solution by systematically identifying inconsistencies, redundancies, and gaps within legal texts. By applying LLMs to municipal laws, townships could proactively assess their legal frameworks, improving clarity, enforcement, and financial sustainability. However, implementing such an approach requires careful consideration of computational constraints, document formats, and the broader applicability of AI-driven legal analysis.

1.3 Thesis Statement

An LLM-based tool to convert a document into an attributed knowledge graph can be used to check for consistency and completeness will allow municipal lawyers to create consistent and complete law documents which prevent costly disputes and reduce revenue losses.

The application of LLMs in legal document analysis has the potential

to revolutionize municipal law by providing an automated, systematic approach to ensuring consistency and completeness. Traditional legal drafting and review processes rely heavily on human oversight, which is inherently susceptible to errors, inconsistencies, and omissions—particularly in laws that have evolved over time through multiple amendments and contributors. By leveraging an LLM-based tool to convert legal documents into attributed knowledge graphs, municipalities can proactively identify gaps, redundancies, and contradictions before laws are enacted or enforced. This proactive approach minimizes ambiguity, strengthens legal clarity, and enhances the efficiency of legal review processes.

A knowledge graph-based representation of municipal laws enables a structured, machine-readable format that facilitates logical analysis. Unlike traditional text-based legal review, which requires extensive manual effort to trace dependencies and resolve conflicts, a knowledge graph explicitly maps relationships between legal provisions, definitions, and enforcement mechanisms. This allows municipal lawyers to assess the interconnectivity of legal clauses and verify their consistency against established legal principles and precedents. Furthermore, an attributed knowledge graph can highlight areas where laws are incomplete or misaligned with overarching governance policies, enabling timely revisions that improve legal coherence.

Beyond legal clarity, the ability to create consistent and complete municipal laws has direct financial implications. Inconsistent or incomplete regulations can lead to disputes over zoning, taxation, permitting, and compliance, often resulting in costly litigation or lost revenue due to unenforceable provisions. By employing an LLM-driven tool to detect and resolve these issues at the drafting stage, municipalities can reduce legal ambiguities that might otherwise be exploited, streamline enforcement mechanisms,

and enhance overall regulatory efficiency. This, in turn, strengthens fiscal sustainability by preventing revenue leakage and ensuring that all applicable fees, fines, and taxes are properly assessed and collected.

The integration of LLM-based tools in municipal lawmaking represents a transformative step toward modernizing local governance. As artificial intelligence continues to advance, municipalities that adopt such technologies will gain a significant advantage in maintaining a legally sound, financially sustainable framework. Future research can extend this approach beyond municipal laws to other domains of legal and regulatory governance, demonstrating the broader impact of AI-driven knowledge representation in ensuring legal accuracy, reducing administrative burdens, and enhancing public trust in local government operations.

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.

The created knowledge graph will be analyzed to determine whether it is appropriate to check the document for inconsistencies and incompleteness. This will include introducing issues in the source documents and then highlighting how easy they are to observe in the knowledge graph.

1.5 Research Questions

To achieve the research objectives, the following research questions will be considered.

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?

1.6 Research Hypotheses

Research will be conducted to test the following hypotheses.

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.

1.7 Research Scope and Limitations

The subsequent sections outline the scope and limitations of this study, which employs Pennsylvania township laws as a case study to develop and evaluate an automated tool for the analysis of legal documents. These publicly accessible laws, having undergone extensive manual reviews for consistency and completeness, provide a rigorous benchmark for assessing the proposed methodology. The primary focus of this research is the

construction of Knowledge Graphs that faithfully represent the structure and content of the documents, thereby laying the groundwork for future efforts in verifying legal consistency and completeness. Notwithstanding, this study acknowledges several inherent limitations, including computational constraints, challenges associated with specific document formats and linguistic nuances, and the primary emphasis on textual analysis. These limitations underscore the necessity for continued research to enhance and broaden the applicability of the proposed approach.

1.7.1 Research Scope

This study focuses on the use of Pennsylvania township laws as a case study for developing and testing an automated tool designed to analyze legal documents. These laws, which are publicly available in both PDF and Word formats, were selected due to their complexity, extensive length, and the fact that they have been authored by multiple contributors over time. Additionally, they have undergone rigorous manual reviews for consistency and completeness, making them an ideal benchmark for evaluating the effectiveness of the proposed approach. While the primary application is in the legal domain, the methodology is designed to be adaptable for broader use across various document types.

The core development in this research centers on constructing Knowledge Graphs that accurately represent the structure and content of the documents under review. These graphs will serve as a foundation for future work in verifying legal consistency and completeness. While the study will assess the suitability of the generated Knowledge Graphs for such tasks, the actual implementation of automated consistency and completeness checks will be left for future research. This approach ensures a focused and systematic

exploration of Knowledge Graph generation while laying the groundwork for subsequent advancements in automated legal analysis.

1.7.2 Research Limitations

This study has several potential limitations. Computational constraints may affect the efficiency and scalability of processing large and complex legal documents. Challenges may also arise in handling specific document formats and language intricacies, particularly in ensuring accurate interpretation and structuring of legal text. Additionally, while this research focuses on leveraging LLMs such as Gemini and ChatGPT, it does not develop specialized models tailored for knowledge graph construction, consolidation, or query answering—an approach that could reduce computational costs and energy consumption.

This research does not perform direct testing for consistency and completeness. Instead, it utilizes Pennsylvania township laws, which are publicly available and have already undergone such validation. Future studies should explore the applicability of this approach to a broader range of legal and non-legal documents.

For document handling, this research mainly uses Word documents to facilitate modifications during testing. Although the methodology should also be compatible with PDFs, further research is needed to confirm seamless integration and processing across different formats.

Finally, this study is limited to textual analysis. Future research could expand upon this work by incorporating additional elements such as tables, formulas, images, and diagrams to improve document comprehension and analysis.

1.8 Praxis Organization

The remainder of this research is organized into several key chapters. Chapter 2 provides a comprehensive review of the relevant literature, focusing on the creation of knowledge graphs from documents by LLMs, the processing of multiple knowledge graphs into a combined knowledge graph by LLMs, the utility of knowledge graphs in representing the original document to ensure consistency and completeness, the process of creating and maintaining local laws in Pennsylvania, and background information on checking documents for consistency and completeness. Chapter 3 delves into the statistical and machine learning methodologies employed in this research, detailing the processes of data pre-processing, model selection, training, and evaluation. Chapter 4 presents and analyzes the results of the data analysis, addressing each research question and hypothesis while evaluating the performance of the proposed methodology and tool. Finally, Chapter 5 concludes the investigation with a discussion of the key findings, contributions to the field, recommendations for practical applications, and potential avenues for future research.