

Document-Level Anomaly Detection in Semi-Structured Text using Large Language Models

E. Patricia Becerra-Sánchez*

Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg

Radu State

Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg

Abstract

Detecting anomalies in semi-structured documents remains a challenging problem due to the complex interaction between textual content, numerical values, and document layout structures. Documents such as financial records, invoices, reports, and administrative forms combine structured and unstructured information, making automated anomaly detection particularly difficult. Traditional detection approaches typically rely on rule-based validation, statistical outlier detection, or classical machine learning models operating on structured representations extracted from document processing pipelines. However, these methods often fail to capture semantic inconsistencies, contextual relationships, and cross-field dependencies embedded in textual document content.

Recent advances in Natural Language Processing (NLP) and large language models (LLMs) provide new opportunities for document-level reasoning over complex textual records. By leveraging contextual language understanding and semantic representation learning, LLMs can integrate textual interpretation with numerical and structural reasoning within a unified document analysis framework. This work investigates how LLM-based architectures can support anomaly detection in semi-structured text documents by enabling contextual understanding of textual descriptions, numerical relationships, and dependencies across document fields.

This work provides three main contributions. First, we present a structured analysis of existing anomaly detection techniques applied to semi-structured documents. Second, we introduce a conceptual taxonomy that organizes rule-based systems, statistical models, deep learning document architectures, and emerging LLM-driven approaches. Third, we analyze how modern NLP architectures—including retrieval-augmented generation and tool-augmented reasoning frameworks—enable semantic inconsistency detection, numerical validation, and contextual reasoning across document sections. The study also highlights key challenges for reliable deployment of LLM-based anomaly detection systems, including numerical reasoning limitations, hallucination risks, robustness in high-stakes environments, and data privacy constraints.

Index Terms

Anomaly Detection, Large Language Models, Natural Language Processing, Semi-Structured Documents