

CHAPTER 1

INTRODUCTION

This chapter discusses the initial foundation of research which includes the background of the problem, the formulation of the problem, the purpose of the research, the benefits of the research, and the limitations of the problem. The discussion in this chapter is directed to show the urgency of research on the implementation of Structure-Aware Hybrid Chunking in the Small Language Model-based Retrieval-Augmented Generation architecture in the Surabaya City regional tax regulation document. Through this chapter, the context of the problem, the direction of the research taken, and the scope of the study are explained so that the research has a clear and measurable focus.

1.1 Background

Regional tax regulations in the city of Surabaya are spread across many Regional Regulations, Mayor's Regulations, and Mayor's Decrees that are interrelated and updated according to policy needs. Surabaya as a metropolitan city accommodates diverse business activities such as hotels, restaurants, entertainment, parking, and other service sectors, each of which carries different terminology, objects, and administrative provisions. These documents are written in a normative, referential, hierarchical style. It is not enough to search for legal information on such documents just by finding keyword matching; Hierarchical contexts such as chapters, articles, verses, dictums, and validity periods are the determinants of the entire meaning. This condition of fragmentation and structural complexity is reflected in the enactment of Surabaya City Regional Regulation (Perda) Number 7 of 2023 concerning Regional Taxes and Regional Levies, which consolidates various types of taxes into one legal umbrella with 104 articles, but still breaks down its operational rules such as tariff classification, determination of the basis for tax imposition, and the provision of incentives into dozens of Mayor Regulations and Mayor's Decrees derivatives [1] .

When the search system for relevant rules is not able to present the structural context in its entirety, the risk of information asymmetry burdens taxpayers and

local governments. The limitations of conventional search systems often fail to maintain the relationship between articles or dictums, so users are vulnerable to misunderstanding obligations, submitting reports late, or misreading certain policy terms. For example, in the Decree of the Mayor of Surabaya regarding the elimination of administrative sanctions, the provisions of fines and interest, the type of tax covered, the subject of the recipient of the facility, and the period of validity are placed in different dictums; A system that presents information partially will make users miss crucial clauses that can change the total policy implications for taxpayers. As a result, municipal governments bear the increasing operational burden of information services, while the failure of the system in providing fast and precise referrals triggers compliance delays that risk lowering the realization of regional revenues. This is very critical considering the dependence of the Surabaya City Regional Revenue and Expenditure Budget (APBD) on very high taxes; in 2024 alone, the target for one sector of the Land and Building Tax (PBB) is pegged at IDR 1.6 trillion, where regulatory confusion often leads to delays in payments or the accumulation of tax receivables that disrupt the liquidity of urban development [2] .

Retrieval-Augmented Generation (RAG) architecture is relevant for these conditions because answers are built from evidence drawn from the corpus, not from the model's "memory" alone. Studies in the legal realm place RAG as an approach to increase interpretability and suppress hallucinations by keeping answers based on source documents, so that article references can be [3] audited. Small Language Models (SLMs) such as Phi-3.5 were chosen because they were more realistically run on limited resources, but the quality of the system relied heavily on how clean and relevant the context was successfully captured; long, noisy contexts tend to undermine the accuracy of answers, especially in high-value scenarios [3][4] .

Regulatory documents are not linear. Structures such as Chapters, Articles, Verse, Letters, or dictums in the Mayor's Decree limit the units of norms that must be understood in their entirety. One article can define a term, another article establishes an object, another article determines sanctions, and that connection is often a prerequisite for reading the provisions correctly. Evaluation studies of

chunking on legal texts show that the way documents are cropped directly affects the quality of retrieval in the RAG system; chunks that are not aligned with the structure can degrade the relevance of the chunk to legal questions [5].

The chunking stage in RAG often starts with fixed-size splitting because it is simple, but in legal documents it risks cutting off the article or dictum in the middle of a sentence so that the evidence taken becomes partial. Semantic chunking attempts to form coherent chunks, but systematic evaluation shows that its computational costs can increase without consistent performance gains on various retrieval and answer generation tasks. Other findings affirm document segmentation as a critical point; [7] Chunks that are too large carry irrelevant information, chunks too small lose coherence, and commonly used rule-based methods often result in suboptimal splits. [8]

The latest literature has already given a strong foothold, especially in two directions. The first direction places segmentation and chunking as the main variables of RAG performance, both through evaluation of the legal corpus and proposals for more adaptive segmentation strategies [5][8]. The second direction addresses retrieval optimization and evaluation for legal domains that demand source-based answers, for example through hybrid retrieval and metrics that examine contextual precision and fidelity on questions of different levels of difficulty [4].

Based on this footing, this study proposes Structure-Aware Hybrid Chunking for the regional tax regulatory corpus of Surabaya City. Truncation is prioritized at the boundaries of a formal structure using regex to keep each chunk corresponding to a reasonable legal unit, in line with the recursive splitting approach that was also evaluated in legal chunking studies [5]. Units that are too long to be contextually retrieval are semantically broken down as a fallback mechanism; this decision is positioned as a compromise between coherence and cost, given that full semantic chunking does not always provide consistent gains [7]. Interarticle dependencies are handled through hierarchical metadata and parent-child relationships, while the context assigned to the model can be elevated to the parent unit when needed, in line with the idea of context reconstruction discussed in the advanced chunking study for RAG [9].

The empirical test was carried out on the Surabaya regional tax regulation corpus, with a dataset of questions that were compiled and labeled as the source of correct article or dictum references. The proposed method is compared to fixed-size chunking as a baseline and semantic chunking as a comparison, as both represent the most common practices on the RAG pipeline [7][8]. Retrieval evaluations use Precision@5, Recall@5, and F1-scores to measure whether correct evidence appears at the top of the list. The accuracy of citations is assessed through evaluation with a manual assessment scheme based on the suitability of answers to the evidence taken, according to the emphasis of RAG evaluation on the legal domain. Computational efficiency is measured through [3][4]. Preprocessing and indexing time, query latency, and the number of chunks generated.

From the series of operational urgency and technical analysis, this study is directed to answer three questions. The first question is how to design chunking that respects the regulatory hierarchy. The second question is how significantly the improvement in the quality of retrieval and accuracy of legal referrals is achieved compared to conventional cutting. The third question is how to compare its computational efficiency when implemented on SLM Phi-3.5 based RAG pipelines. The answers to these three questions are expected to be the basis for a question and answer system for regional tax regulations that can be audited through article references and remain realistically carried out on the computing resources available in the local government environment.

1.2 Problem Formulation

Based on the background that has been described, the formulation of the problem in this study is as follows:

1. How is the implementation of Structure-Aware Hybrid Chunking in the Small Language Model-based Retrieval-Augmented Generation system for Surabaya City Tax Regulation documents?
2. How does Structure-Aware Hybrid Chunking affect the retrieval quality of the system being built?
3. How does Structure-Aware Hybrid Chunking affect the accuracy of the answer to the two-hop reasoning question, the quality of the answer text

based on ROUGE, and the computational efficiency based on the chunking processing time and retrieval time?

1.3 Research Objectives

Based on the formulation of the problem that has been described, the research objectives in this study are as follows:

1. To design and implement Structure-Aware Hybrid Chunking on a Small Language Model-based Retrieval-Augmented Generation system for the processing of Surabaya City Tax Regulation documents that have a hierarchical structure.
2. To measure and analyze the effect of Structure-Aware Hybrid Chunking on retrieval quality in the built system.
3. To test and analyze the effect of Structure-Aware Hybrid Chunking on the accuracy of the answer to the two-hop reasoning question, the quality of the answer text based on ROUGE and the computational efficiency based on the chunking processing time and retrieval time.

1.4 Research Benefits

This research is expected to provide benefits both theoretically and practically for various related parties, as described in the following sub-chapter.

1.4.1 Theoretical benefits

This research contributes to the study of Natural Language Processing and Information Retrieval through a measured evaluation of chunking techniques in hierarchical legal documents. The focus of the contribution lies in the relationship between regulatory structure-based segmentation and retrieval performance and response quality in the Retrieval-Augmented Generation pipeline.

1.4.2 Practical Benefits

Practically, this research is expected to provide benefits for various parties as follows.

- a. For the Surabaya City Government: this research has the potential to be the basis for the design of a prototype tax regulation question and answer

assistant that is able to display article references clearly, so that access to information is faster and easier to trace in the Surabaya City Tax Regulation.

- b. For System Developers: this study provides an empirical comparison of chunking strategies in Retrieval-Augmented Generation architecture for legal domains, thus aiding in the selection of the most appropriate chunking configuration based on measurable metrics.
- c. For Other Researchers: this study can be an experimental baseline and provide a dataset of questions and answers along with source references in the Surabaya City Tax Regulation for testing retrieval and generation methods in the context of regional regulations.

1.5 Problem Limitations

The limitations of the problem in this study are as follows.

1. The research data uses the Surabaya City Regional Regulation document on Regional Taxes and Regional Levies which is still valid as the main corpus of research.
2. The documents used are digital documents with text that can be extracted directly to support the preprocessing, chunking, retrieval, and evaluation processes.
3. The scope of the research material includes the provisions of regional taxation contained in the Regional Regulation, including parts that have the structure of Chapters, Articles, Verses, Letters.
4. The system is built using the Retrieval-Augmented Generation architecture with three main components, namely document preprocessing, chunking, and retrieval-generation.
5. The main method implemented was Structure-Aware Hybrid Chunking, with fixed-size chunking and semantic chunking used as comparative methods at the evaluation stage.
6. The language model used as the generator is an instruct/mini version of Phi-3.5 running on the Kaggle computing environment.
7. The retrieval method used is hybrid retrieval which combines dense embedding retrieval and sparse retrieval based on BM25 with a top-k of 5.

8. The evaluation dataset consists of factual questions, comparative or conditional questions, and computational questions prepared based on the content of the Regional Regulation document.
9. The complexity of the questions is focused on single-hop and two-hop reasoning in the same document.
10. Evaluation of retrieval performance using Precision@5, Recall@5, and F1-score metrics.
11. Evaluation of the quality of the answers using the accuracy of two-hop reasoning, ROUGE-1, ROUGE-2, and ROUGE-L.
12. Evaluation of computational efficiency using chunking processing time and retrieval time in the Kaggle experimental environment.
13. The results of the research were realized in the form of a computational prototype to test the implementation of the method and its performance in the corpus of regional tax regulations of the City of Surabaya.