

CHAPTER V

CONCLUSIONS AND RECOMMENDATION

This chapter presents research conclusions and development suggestions based on the results of implementation, testing, and discussion that have been described in Chapter IV. The conclusions were prepared by referring to the formulation of the problem and the objectives of the research, especially regarding the implementation of Structure-Aware Hybrid Chunking, its effect on retrieval quality, its effect on answer generation, and the efficiency of corpus updates in the Retrieval-Augmented Generation pipeline. Suggestions were given as a direction for further development so that the system can be improved in terms of document representation, evidence selection, generator model, index efficiency, and application prototype.

5.1 Conclusion

Based on the results of experiments, data analysis, and the implementation of the Retrieval-Augmented Generation (RAG) prototype in the Surabaya City regional tax regulation document, several conclusions were drawn as follows:

1. The implementation of Structure-Aware Hybrid Chunking (SAHC) on the Small Language Model (SLM)-based Retrieval-Augmented Generation (RAG) system for Surabaya City Tax Regulation documents was successfully carried out by placing the boundaries of formal legal structures (such as Chapters, Articles, Verses, and Letters) as the main deduction reference using regex. If a legal unit exceeds the maximum limit of a token, a semantic fallback mechanism (fallback) is used so that the unit does not lose its coherence. Through this approach, each chunk is able to bind its hierarchical metadata in its entirety, so that the evidence pulled by the system always carries a normative identity (source of article/verse) that is ready to be processed by SLM such as Phi-3.5 and Qwen 2.5.
2. The influence of SAHC on the quality of retrieval has been proven to be very significant in improving the system's ability to find the right evidence. The test results showed that the combination of SAHC with Hybrid

Reciprocal Rank Fusion (RRF) achieved the best performance compared to conventional methods, with a Precision@5 value of 0.3638, Recall@5 of 0.7632, and F1@5 of 0.4596. The high value of the recall indicates that the formation of chunking based on the regulatory structure effectively prevents the missing of reference documents (gold evidence), and successfully reduces the failure rate (document miss) to the lowest point (45 out of 190 questions).

3. The influence of SAHC on answer generation and system efficiency showed varied but measurable results. On the two-hop reasoning question, SAHC proved its resilience in maintaining the relationship between articles with F1@5 which remained superior at 0.4215. The quality of the answer text based on the ROUGE lexical score was low (ranging from 0.3301 at Qwen 2.5 and 0.1193 at Phi-3.5), but this was due to the tendency of generative models to paraphrase legal language, not because the system took the wrong legal base. In terms of efficiency, the structure-conscious chunking process does require a slightly higher parsing time at the beginning, but the index update mechanism is highly efficient through incremental dense re-indexing with a speedup of 47,719 times faster than a full rebuild computational process.

Table 5.1 Summary of the main conclusions of the study

Aspects	Key findings	Significance of research
Retrieval	SAHC + hybrid RRF memperoleh Precision@5 0,3638, Recall@5 0,7632, dan F1@5 0,4596.	The legal structure maintained in the chunk has been proven to improve the quality of evidence search.
Top-k and hop	SAHC+ hybrid RRF continues to excel in P@1, P@3, single-hop, and two-hop.	Relevant evidence appears more often in the early stages and is more readily available for use in RAG.
Answer generation	Local Phi-3.5 is not yet stable, while extractive baseline and Qwen 2.5 are more feasible for prototypes.	The quality of the generator needs to be distinguished from the quality of retrieval; Evidence selection remains key.

Aspects	Key findings	Significance of research
Re-indexing	Incremental dense re-indexing is more suitable for partial updates, whereas BM25 still requires a full rebuild.	The pipeline is more realistic for a corpus of regulations that can change.
Purwarupa	Streamlit contains the artifacts of the experiment and displays answers, evidence, metadata, and evaluations.	The results of the research can be observed in the form of auditable question and answer applications.

Table 5.1 summarizes that the main contribution of the study is on the quality of retrieval and traceability of evidence. Answer generation and prototype testing showed that better evidence needed to be followed by context control and appropriate generator selection. These results make SAHC's position clearer as a retrieval foundation in the RAG system of regulatory documents, rather than as a replacement for the entire generator component.

5.2 Suggestions

In order to improve the constraints and limitations of the system found throughout the research, several suggestions for further development were proposed as follows:

1. In the design and implementation of the pre-processing of chunking hierarchical legal documents in the future, it is recommended to strengthen the structure parser to be more adaptive in handling documents outside the Regional Regulation, such as the Mayor's Decree which uses a lot of dictum, complex tariff table attachments, and decision items. In addition, the Layout-Aware OCR integration is highly recommended for handling scanned regulatory documents so that hierarchical numbering is not damaged before chunking.
2. To overcome the low precision value due to noise from homogeneous tax document redacting, retrieval quality testing is further recommended to include an advanced screening method (structure-based Context Re-ranking) after the Top-5 withdrawal process. In addition, further research can apply the parent-child retrieval mechanism, which is to look for specific pieces (verses/letters) for precision, then pull the parent unit (whole article) to be presented to the generator so that the legal context is not interrupted.

3. Regarding testing the performance of reasoning and the quality of the text, evaluation using ROUGE should be accompanied by a legal expert's manual assessment or based on LLM-as-a-Judge, considering that the correctness of the regulatory text cannot be judged purely from the similarity of words (lexical). It is also necessary to fine-tune based on legal data (legal domain tuning) in the local SLM so that the answer format and how to refer to the article become more consistent. In terms of the efficiency of corpus updates, the incremental indexing system should be equipped with regulatory versioning to handle tax rules that are revoked or changed without manual index computing (full rebuild).