



UNDERGRADUATE THESIS

Implementation of Structure-Aware Hybrid Chunking in Retrieval-Augmented Generation (RAG) for Small Language Model (SLM) in Surabaya City Regional Tax Regulation Documents

ACHMAD YUSUF YULESTIONO
NPM 22081010180

THESIS ADVISORS

Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.
Ardhon Rakhmadi, S.Tr.T., M.Kom

**MINISTRY OF HIGHER EDUCATION, SCIENCE, AND TECHNOLOGY
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FACULTY OF COMPUTER SCIENCE
INFORMATICS STUDY PROGRAM
SURABAYA
2026**

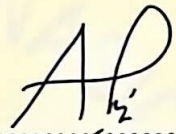
Implementation of Structure-Aware Hybrid Chunking in Retrieval-Augmented Generation (RAG) for Small Language Model (SLM) in Surabaya City Regional Tax Regulation Documents

By:
ACHMAD YUSUF YULESTIONO
NPM. 22081010180

Has been defended before, and accepted by, the Board of Assessors of the Thesis Examination of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional Jawa Timur, on June 24, 2026:

Approved,

Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT..
NPT. 222198 60 816400


.....

(Advisor I)

Ardhon Rakhmadi, S.Tr.T., M.Kom.
NIP. 19910805 202406 1 002


.....

(Advisor II)

Dr. Intan Yuniar Purbasari, S.Kom, M.Sc.
NIP. 19800602 202521 2 029


.....

(Head Assessor)

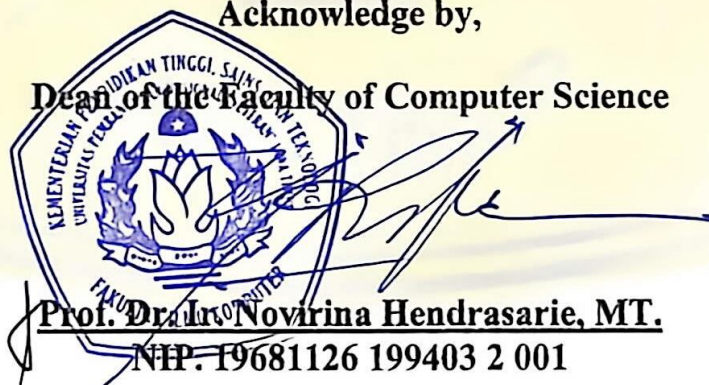
Hazna At Thooriqoh, S.Tr.Kom. M.Kom.
NIP. 19970320 202406 2 003


.....

(Assessor I)

Acknowledge by,

Dean of the Faculty of Computer Science


.....

Prof. Dr. Ir. Novirina Hendrasarie, MT.
NIP. 19681126 199403 2 001

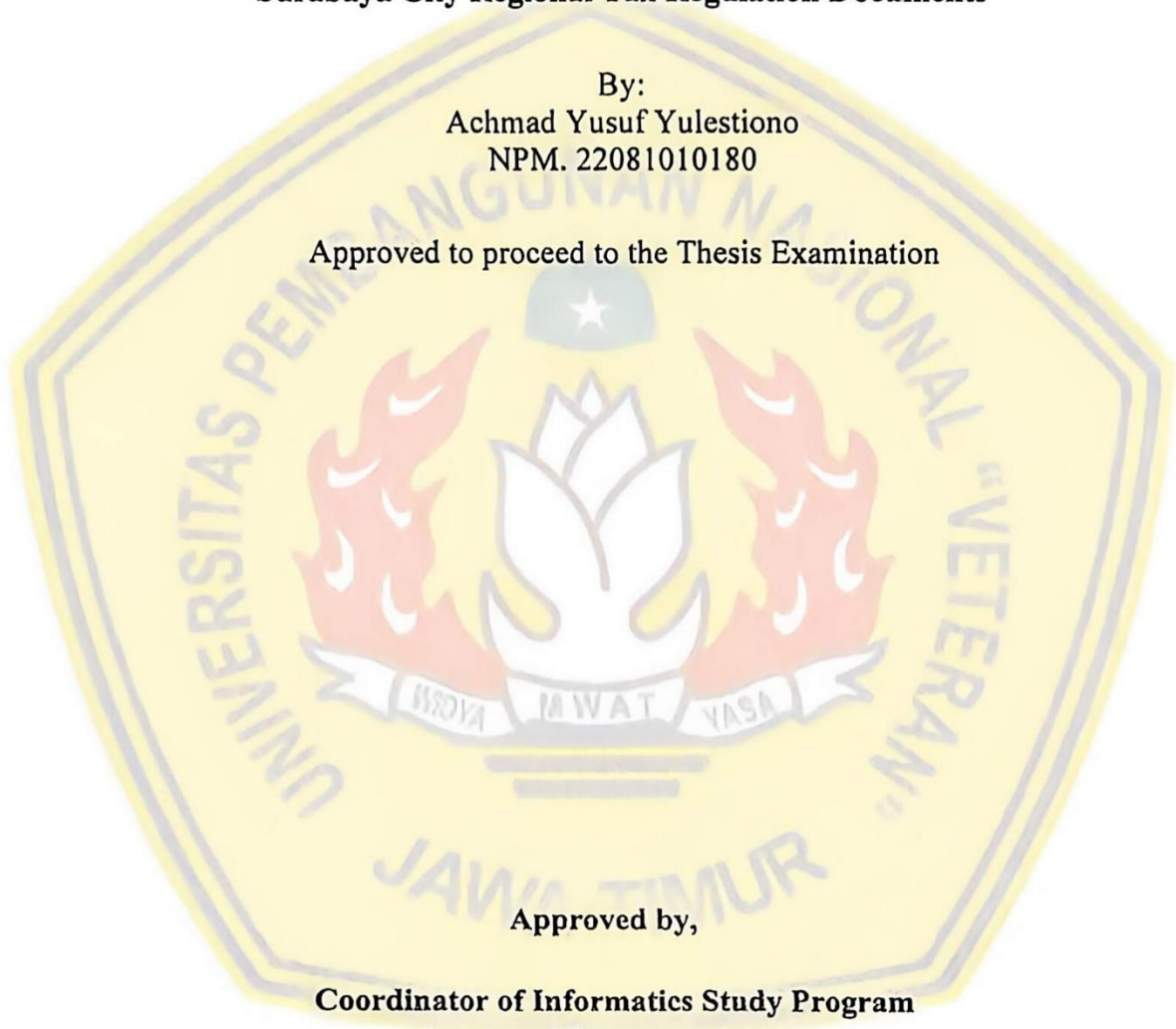
APPROVAL SHEET

Implementation of Structure-Aware Hybrid Chunking in Retrieval-Augmented Generation (RAG) for Small Language Model (SLM) in Surabaya City Regional Tax Regulation Documents

By:

Achmad Yusuf Yulestiono
NPM. 22081010180

Approved to proceed to the Thesis Examination



Approved by,

**Coordinator of Informatics Study Program
Faculty of Computer Science**

Dr. Intan Yuniar Purbasari, S.Kom. MSc.
NIP. 19800602 202521 2 029

STATEMENT OF ORIGINALITY

The undersigned:

Student Name : Achmad Yusuf Yulestiono

Study Programs : Computer Science

Supervisor : 1. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.
2. Ardhon Rakhmadi, S.Tr.T., M.Kom

It is hereby stated that the contents of part or all of the thesis with the title:

Implementation of Structure-Aware Hybrid Chunking in Retrieval-Augmented Generation (RAG) for Small Language Model (SLM) in Surabaya City Regional Tax Regulation Documents

is truly the work of an independent intellectual, completed without the use of unauthorized materials and is not the work of another party that I acknowledge as my own work. All references cited or referenced have been written in full in the bibliography. If it turns out that this statement is not true, I am willing to accept sanctions in accordance with the applicable regulations.



Surabaya, June 24, 2026

Author,



ACHMAD YUSUF YULESTIONO
NPM. 22081010180

ABSTRACT

Student Name / NPM : Achmad Yusuf Yulestiono / 22081010180
Thesis Title : Implementation of Structure-Aware Hybrid
Chunking in Retrieval-Augmented Generation
(RAG) for Small Language Model (SLM) in
Surabaya City Regional Tax Regulation
Documents

Supervisor : 1. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.
2. Ardhon Rakhmadi, S.Tr.T., M.Kom

Local tax regulatory documents have a complex hierarchy that makes it difficult to find keyword-based information. This study implements Structure-Aware Hybrid Chunking (SAHC) on the Retrieval-Augmented Generation (RAG) architecture to break down documents based on the boundaries of legal structures (Chapters, Articles, Verses) and semantics, using the Small Language Model Phi-3.5. The performance of SAHC was compared to fixed-size and semantic chunking through three test scenarios. The results showed that SAHC with hybrid Reciprocal Rank Fusion (RRF) achieved the best retrieval performance with Precision@5 (0.3638), Recall@5 (0.7632), and F1@5 (0.4596). The high quality of evidence search is able to support the accuracy of answers, although the stability of text generation requires strict evidence screening. In addition, corpus updates have proven to be more efficient using incremental dense indexing. In conclusion, SAHC proves to be an optimal segmentation approach for hierarchical documents because it is able to maintain structural integrity and ensure that answers can be traced back to their original normative sources.

Keywords: Retrieval-Augmented Generation, Structure-Aware Hybrid Chunking, Small Language Model, regulasi pajak daerah, hybrid retrieval, Reciprocal Rank Fusion

ACKNOWLEDGEMENTS

Praise be to Allah SWT for all His graces, guidance, and gifts to the author so that the thesis with the title "**Implementation of Structure-Aware Hybrid Chunking in Retrieval-Augmented Generation (RAG) for Small Language Model (SLM) in Surabaya City Regional Tax Regulation Documents**" can be completed properly. The author realizes that the completion of this thesis would not have been possible without the continuous support, guidance, and assistance from various parties. Therefore, the author would like to express profound gratitude to::

1. Allah SWT, for every blessing, every obstacle turned into ease, and every moment of strength in times of uncertainty. None of this journey would have been possible without His will and guidance.
2. The author's beloved parents, whose endless prayers, sacrifices, patience, and unwavering support have been the foundation of every achievement. Their unconditional love and belief have been the greatest source of strength throughout this academic journey.
3. Mrs. Prof. Dr. Ir. Novirina Hendrasarie, M.T. as the Dean of the Faculty of Computer Science, National Development University "Veteran" East Java.
4. Mrs. Dr. Intan Yuniar Purbasari, S.Kom. M.Sc. as the Coordinator of Informatics Study Program, Faculty of Computer Science, National Development University of East Java.
5. Mrs. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT., as Supervisor I, and Mr. Ardhon Rakhmadi, S.Tr.T., M.Kom., as Supervisor II. The author is deeply grateful for their patience, guidance, encouragement, and invaluable advice throughout the preparation of this thesis. Their dedication and willingness to share knowledge have greatly contributed to the completion of this research.
6. Mrs. Dr. Intan Yuniar Purbasari, S.Kom., M.Sc., as Examiner I, and Mrs. Hazna At Thooriqoh, S.Tr.Kom., M.Kom., as Examiner II, for their constructive feedback, valuable suggestions, and insightful evaluations that helped improve the quality of this thesis

7. Lecturers of the Informatics Department, National Development University of East Java Veterans for their guidance during the author's education.
8. Muhammaf Reyhan Dwi Fisena, as a close friend who accompanied the author through various challenges and moments of growth. The author truly appreciates his continuous support, valuable advice, and companionship. His encouragement and presence have been invaluable during the completion of this thesis.
9. My cancer, which has colored the process of writing this thesis in ways the author never anticipated. Through the pain and the heavy days, it became an unexpected teacher of gratitude, endurance, and hope. It forced the author to slow down, yet it also gave the author the profound courage to keep moving forward and prove that adversity does not define the end of a journey.
10. All friends and colleagues in the Informatics Study Program, for the shared experience and solidarity throughout this academic journey. Special recognition and deepest gratitude are also extended to Rangga Agni, Pramudya Rekhsa, Rafie Najwa, and Alief Listanto, for their exceptional support, endless motivation, and meaningful presence that made this entire process much more bearable and memorable.
11. The author would also like to express sincere gratitude to Kevin Ricky, Gamar Ramadhani, Mario Subagio, and Mulyani Bhakti for their unwavering support and dedicated assistance throughout the entirety of this research. Their practical help, continuous encouragement, and willingness to assist the author (and sorry for being silly at times!) during the challenging phases of writing were instrumental in bringing this thesis to its successful completion.
12. To my dearest friends, Ludiana Syahid and Adin Fath, words cannot express how profoundly grateful the author is for your presence. Through the countless sleepless nights and overwhelming challenges, your genuine empathy, comforting words, and silent understanding have kept the author moving forward. Thank you for holding my hand through the darkest days and for reminding me of my own strength when I had forgotten it. This milestone is as much yours as it is mine.

13. To the unseen soul whose name has long been etched in the sacred ink of Lauhul Mahfudz. Though my eyes have yet to find you, my heart has felt your silent strength, pulling me through the darkest and heaviest of my days. I have built a fortress around this heart, fiercely resisting the hollow warmth of passing infatuations even when they were within my grasp. You have been a distant beacon of hope, an unspoken motivation when my own spirit wavered. If the heavens have decreed this transient earth as the place our paths finally intertwine, I pray our convergence becomes a sanctuary of endless blessings. May we walk this fleeting world hand in hand a journey woven in grace until our steps lead us safely home to the eternal dawn of Jannah.

14. And finally, a confession to my own reflection: I owe you the greatest apology of all. Forgive me for being your most unforgiving tormentor, for treating your tender heart with such unwarranted cruelty. I am profoundly sorry for turning a blind eye to your quiet weeping, constantly invalidating the depths of your pain. Forgive me for locking your vibrant spirit in a cage of my own making, clipping your wings just to fit into spaces meant for others. Most of all, I am sorry for the countless times I swallowed your truth, dimming your light and shrinking into the shadows merely to keep everyone else warm. You deserved a far gentler guardian. Thank you for enduring me, and for surviving it all.

The author realizes that in the preparation of the following thesis there are many shortcomings. For this reason, constructive criticism and suggestions from all parties are highly expected for the perfection of writing the following thesis. Finally, with all the limitations that the author has, hopefully the following report can be useful for all parties in general and the author in particular.

Surabaya, June 24th 2026

Author

ACHMAD YUSUF YULESTIONO

NPM. 22081010180

TABLE OF CONTENTS

APPROVAL SHEET	ii
APPROVAL SHEET	iii
STATEMENT OF ORIGINALITY	iv
ABSTRACT	v
ACKNOWLEDGEMENTS.....	vi
TABLE OF CONTENTS.....	ix
LIST OF FIGURES	xiii
LIST OF TABLES	xv
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.2 Problem Formulation	4
1.3 Research Objectives	5
1.4 Research Benefits.....	5
1.4.1 Theoretical benefits	5
1.4.2 Practical Benefits.....	5
1.5 Problem Limitations.....	6
CHAPTER 2 LITERATURE REVIEW	8
2.1 Previous Research	8
2.2 Characteristics of Regional Tax Law Regulation Documents as Objects of Study 13	
2.2.1 Hierarchical Structure of Chapters, Chapters, Verses and Letters... 13	
2.2.2 Definitions, Operational Conditions, and Cross-References	15
2.2.3 Implications of Normative Structure on Augmented Generation Retrieval	15
2.3 System Paradigm: Retrieval-Augmented Generation for Legal Domains 18	
2.4 Small Language Model on the RAG Pipeline and Context Design Implications.....	21
2.5 Text Segmentation and Chunking.....	25
2.5.1 The Role of Chunking in the RAG Pipeline	25

2.5.2 Fixed-Size Chunking.....	26
2.5.3 Semantic Chunking	27
2.5.4 Structure-Aware Chunking	29
2.5.5 Structure-Aware Hybrid Chunking	30
2.6 Hybrid Retrieval on Legal Documents	32
2.7 Ground Truth, Evidence, and Question Dataset.....	35
2.7.1 Evidence on legal QA	36
2.7.2 Single-hop and two-hop questions	37
2.7.3 Categories of factual, definitional, and conditional questions	38
2.8 Foundations of Evaluation: Retrieval, Two-Hop Reasoning, ROUGE, and Efficiency Metrics	40
2.8.1 Evaluasi Retrieval: Precision@5, Recall@5, dan F1-score	40
2.8.2 Limited Reasoning Evaluation: Two-Hop Reasoning Accuracy	41
2.8.3 Evaluation of the Quality of the ROUGE-1, ROUGE-2, and ROUGE-L Answer Texts	42
2.8.4 Efficiency Evaluation: Chunking Processing Time and Retrieval Time	
43	
CHAPTER 3 METHODOLOGY	46
3.1 Research Design.....	46
3.2 Research Data.....	50
3.2.1 Data Sources.....	50
3.2.2 Query Dataset.....	51
3.2.3 Statistics Dataset	54
3.2.4 Data management and reproducibility	54
3.3 Research System Flow / Architecture	55
3.3.1 Preprocessing Document.....	57
3.3.2 Application of the Chunking Method	61
3.3.2 Data Treatment Sample	64
3.3.4 Embedding and Indexing	69
3.3.5 Retrieval	70
3.3.6 Generation	71

3.4 Prototype Design of Web-Based Systems.....	72
3.5 Research instruments	77
3.5.1 Hardware	77
3.5.2 Software and Libraries	78
3.6 Test Mechanism and Experimental Scenario	80
3.6.1 General Principles of Testing.....	80
3.6.2 Experimental parameters and controlled configurations.....	81
3.6.3 Pilot Phase (Limited Trial).....	83
3.6.4 Scenario I: Retrieval Testing.....	84
3.6.5 Scenario II: RAG End-To-End Testing.....	85
3.6.6 Scenario III: Corps Update Testing and Re-indexin Efficiency	86
3.6.7 Recording of Experimental Outputs.....	90
3.7 Evaluation Methods	90
3.7.1 Evaluation of Retrieval Quality	91
3.7.2 Evaluation of Two-Hop Reasoning.....	91
3.7.3 Generation Quality Evaluation.....	92
3.7.4 Evaluation of Computational Efficiency.....	93
CHAPTER IV RESULTS AND DISCUSSION	95
4.1 Overview of Execution Results.....	95
4.2 Initial System Implementation Results	97
4.2.1 Document Preprocessing Results	97
4.2.2 Parser Metadata Results	99
4.2.3 Hasil Structure Parser.....	100
4.2.4 Results of the Chunking Method Application.....	102
4.2.5 Results of Text Retrieval and Indexing.....	103
4.3 Outcome of Scenario I: Retrieval Testing.....	104
4.3.1 Configuration and Scale of Retrieval Measurements.....	105
4.3.2 Retrieval Metric Results.....	107
4.3.3 Results Based on Reasoning Hop	110
4.3.4 Error Retrieval Analysis.....	112
4.4 Outcome of Scenario II: End-to-End Testing of RAG.....	113

4.4.1 Configuration and Scale of Measurement of Answer Generation .	114
4.4.2 ROUGE Results and Generation Success	116
4.4.3 Sample Output and Conformity of Evidence	118
4.5 Outcome of Scenario III: Corpus Update Testing and Re-indexing Efficiency	120
4.5.1 Corpus Update Configuration	120
4.5.2 Re-indexing Time Results	121
4.5.3 Impact of Corpus Update on Retrieval.....	123
4.5.4 Re-indexing Efficiency Analysis	124
4.6 General discussion	125
4.6.1 Comparison of Results Between Scenarios.....	125
4.6.2 Effect of Structure-Aware Hybrid Chunking on Retrieval	126
4.6.3 The Effect of Structure-Aware Hybrid Chunking on Answer Generation	128
4.6.4 Computational Accuracy and Efficiency Trade-offs	129
4.7 Implications for System Prototypes	131
4.7.1 Position of the Prototype in Research	131
4.7.2 Streamlit Interface Implementation.....	133
4.7.3 Prototype Operational Workflow	135
4.7.4 Evidence Traceability and Evaluation Summary	136
4.7.5 Implications and Limitations of Prototypes	137
CHAPTER V CONCLUSIONS AND RECOMMENDATION.....	139
5.1 Conclusion.....	139
5.2 Suggestions	141
REFERENCES.....	143
APENDIX	147

LIST OF FIGURES

Figure 2. 1 Example of the contents of PERWALI Surabaya number 72 2023 ...	14
Figure 2. 2 Experimental Design Diagrams of 2 levels or more.....	19
Figure 2. 3 Advantages of SLM.....	22
Figure 2. 4 Strategy levels chunking.....	30
Figure 3. 1 Flow of research stages.....	46
Figure 3. 2 Architecture Pipeline RAG end-to-end	56
Figure 3. 3 Flowchart Proses Chunking.....	63
Figure 3. 4 Proses preprocessing data	65
Figure 3. 5 Prototype System Process Diagram.....	73
Figure 3. 6 Wireframe System Interface Design.....	75
Figure 3. 7 Wireframe Interface Design Generation Results Evaluation.....	76
Figure 4. 1 Screenshot of the experiment configuration on the notebook	96
Figure 4. 2 Screenshot of the original output of the PDF preprocessing summary	97
Figure 4. 3 PDF extraction results graph	98
Figure 4. 4 Screenshot of the original output of the parser metadata summary....	99
Figure 4. 5 Screenshot of the original output number of units of the legal structure	101
Figure 4. 6 Graph of the number of units of legal structure.....	102
Figure 4. 7 Screenshot of the original output of the result statistics chunking ...	102
Figure 4. 8 Quantity Chart Chunk and Word Average	103
Figure 4. 9 Summary Output Screenshot Indexing.....	104
Figure 4. 10 Screenshot of the output of the evaluation question summary	106
Figure 4. 11 Screenshot of the original output of the evaluation results retrieval top- 5.....	107
Figure 4. 12 Graph F1@5 per combine retrieval	108
Figure 4. 13 Heatmap F1@5 results retrieval	108
Figure 4. 14 Screenshot of the original output retrieval Based on Reasoning Hop	110

Figure 4. 15 Precision Changes at $K = 1$, $K = 3$, and $K = 5$	111
Figure 4. 16 Screenshot of the error summary output retrieval	112
Figure 4. 17 Jumlah document miss@5	113
Figure 4. 18 Screenshot of the original output of the row-level selection evidence	115
Figure 4. 19 Screenshot of the local Phi-3.5 generation summary output	116
Figure 4. 20 Screenshot of Qwen 2.5 API generation summary output	117
Figure 4. 21 Comparison of ROUGE generation results	117
Figure 4. 22 Screenshot of the original output of the ROUGE generation Qwen 2.5 graph.....	119
Figure 4. 23 Screenshot of the original output log re-indexing fixed-size.....	121
Figure 4. 24 Screenshot of the original output log re-indexing semantic	121
Figure 4. 25 Screenshot of the original output log re-indexing SAHC.....	121
Figure 4. 26 Screenshot of the original output full time and incremental dense re- indexing.....	121
Figure 4. 27 Comparison of full and incremental time dense re-indexing.....	122
Figure 4. 28 Screenshot of the original output impact of the corpus update on Hit@5	123
Figure 4. 29 Screenshot of the output of the new document Hit@5 graph.....	123
Figure 4. 30 The impact of corpus updates on Hit@5	124
Figure 4. 31 Streamlit prototype view.....	133
Figure 4. 32 Evaluation tab view on Streamlit prototype	134
Figure 4. 33 Tab view Evidence Top	136

LIST OF TABLES

Table 2. 1 Previous Research	9
Table 2. 2 Research Metrics and Aspects	44
Table 3. 1 Question Categories	53
Table 3. 2 Example Results Fixed-size Chunking	66
Table 3. 3 Example Results Semantic Chunking	66
Table 3. 4 Example Results Structure-Aware Hybrid Chunking	67
Table 3. 5 Attribute Details Chunking	68
Table 3. 6 Software and Libraries Used	78
Table 3. 7 Test Hyperparameters	82
Table 4. 1 Summary of experimental execution results	96
Table 4. 2 Summary of PDF preprocessing results	98
Table 4. 3 Summary of parser metadata results	100
Table 4. 4 Number of detected structural units	101
Table 4. 5 Results Statistics Chunking	103
Table 4. 6 Summary of results indexing	104
Table 4. 7 Query dataset finalization set for testing retrieval	106
Table 4. 8 Evaluation results retrieval top-5	107
Table 4. 9 Results retrieval Based on Reasoning Hop	110
Table 4. 10 Summary of errors retrieval	112
Table 4. 11 Scenario II testing process series	115
Table 4. 12 Summary of generation results	117
Table 4. 13 Example of output generation for analysis	119
Table 4. 14 Results of the re-indexing	122
Table 4. 15 The impact of corpus updates on Hit@5	123
Table 4. 16 Comparison of key outcomes between scenarios	126
Table 4. 17 The effect of SAHC on quality retrieval	127
Table 4. 18 Influence evidence selection of answer generation	128
Table 4. 19 Trade-offs of computing accuracy and efficiency	130
Table 4. 20 Prototype position of the research stage	132

Table 4. 21 Main components of the prototype interface	134
Table 4. 22 Artifacts used by the prototype	137
Table 4. 23 Summary of the implications of the prototype for research.....	138