Effective Chunking for Retrieval-Augmented Generation over Structured Institutional Documents

Siti Sarah Izhan Khalib[1], Abdul Hadi Abd Rahman[2]*, Lailatul Qadri Zakaria[3]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia[1]
Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia[2, 3]

***Abstract*—Retrieval-Augmented Generation (RAG) has improved domain-specific question answering tasks by grounding responses in an external knowledge base. However, institutional documents differ from general corpora due to the presence of structured content such as tables, flowcharts, fee breakdowns and program codes. Existing RAG pipelines typically employ generic chunking strategies designed for unstructured text, which can fragment structured information and degrade answer quality. This study evaluates four chunking strategies across six configurations. The evaluation uses 65 publicly available structured documents from Universiti Kebangsaan Malaysia (UKM) with 28 Bahasa Melayu queries. Retrieval performance, answer quality and structural integrity are compared across configurations. Retrieval performance is generally similar across configurations, whereas answer quality varies more substantially. Structure-aware chunking achieves the highest ROUGE-L, BLEU and BERTScore F1 values, while semantic chunking achieves the highest table-integrity score. The findings demonstrate that chunking strategies influence RAG objectives in distinct ways, with structure-aware chunking improving answer generation while semantic chunking more effectively preserving table integrity. The results provide an empirical basis for selecting chunking configurations for structured institutional RAG systems.**

***Keywords*—Chunking; retrieval-augmented generation; structured text; documents**

I. Introduction

The rapid advancement of Large Language Models (LLMs) has significantly enhanced the capabilities of intelligent chatbot systems across multiple domains, including higher education [1]-[5]. Universities increasingly deploy chatbot systems to support university communities' inquiries related to students' admissions, academic programs and administrative procedures [5]. However, traditional FAQ-based systems are limited by their reliance on predefined question-answer pairs [6], restricting their ability to handle diverse or unseen queries and reducing their flexibility.

The emergence of Retrieval-Augmented Generation (RAG) for chatbots has addressed these limitations by integrating external knowledge sources into the response generation process. RAG enables chatbots to retrieve relevant information from domain-specific documents, improving factual accuracy and reducing hallucination [7]. In the context of higher education, RAG-based systems have been explored to enhance institutional question answering by leveraging university documents such as guidelines, brochures and policy materials [8].

*Corresponding author.

Despite these advancements, several challenges remain in the retrieval process for institutional question answering. Unlike general unstructured text corpora, institutional documents often contain structured information such as tables, flowcharts, fee breakdowns and program codes [9]. However, standard text extraction pipelines are typically designed for linear and unstructured text, which may fail to preserve these structural elements effectively [9], [10]. As a result, important relationships within tables can be lost, leading to noisy representations and broken structural context even before the chunking stage. This limitation reduces the quality of the knowledge base and introduces errors throughout the retrieval process.

In addition, a more critical challenge lies in the chunking strategy used in RAG systems. Most existing approaches adopt generic chunking methods [7], [11] such as fixed-size or recursive segmentation [12], which are primarily designed for unstructured narrative text. When these methods are applied to structured institutional documents, they may split tabular data or logically related information across multiple chunks [13]. This situation results in fragmented context and degradation of the answer quality [12]. Although prior studies have explored various chunking strategies in RAG [7], [12], relatively limited evidence is available on how these strategies preserve semantic and structural integrity within structured document settings, particularly in real-world institutional environments.

To address this gap, the study evaluates the effect of chunking configurations on RAG performance for structured institutional documents. Four chunking strategies are examined through six configurations: fixed-size 256 tokens, fixed-size 512 tokens, fixed-size 512 tokens with 50-token overlap, recursive, semantic and structure-aware chunking. The configurations are compared in terms of retrieval effectiveness, answer quality and structural integrity preservation. The main contributions of the study are as follows: 1) an empirical comparison of six chunking configurations for structured institutional documents, 2) a structure-sensitive evaluation of table integrity, and 3) an empirical analysis based on a real-world institutional dataset derived from Universiti Kebangsaan Malaysia (UKM).

II. Related Work

RAG systems combine retrieval and generation to ground LLM responses based on external knowledge [7], and recent studies have investigated their use in higher education and other domain-specific settings [8], [12]. In these systems, chunking determines the units of the data that are indexed in the vector databases and retrieved, thereby influencing both the

information available to the generator and the computational cost of retrieval.

Recent chunking research has moved beyond simple fixed-sized segmentation. Recursive methods use hierarchical separators to preserve local context, while semantic methods use embedding similarity to form meaning-based segments [11], [12], [13]. Studies in clinical, enterprise and scientific domains indicate that chunk size and boundary selection can affect retrieval and downgrade the answer quality [12], [14], [15]. These findings suggest that retrieval effectiveness alone may not capture the full impact of chunking on generation.

Chunking becomes more challenging when documents contain structured and heterogeneous content, as separating logically related elements can affect the information available for generation. In such settings, related elements may need to remain together when their relationship is necessary to answer a query. TableRAG, for example, emphasizes the preservation of table structure for heterogeneous document reasoning [16]. Structure-aware chunking has also been investigated for enterprise documents, where document elements can be segmented according to their structural roles [14], [17]. Recent topology-aware approaches further consider hierarchical relationships between document components [18]. However, evidence remains limited for multilingual institutional documents containing heterogeneous PDFs, tables, programme codes and administrative content.

The present study therefore focuses on the relationship between chunk boundaries, retrieval behaviour, answer quality and structural integrity in an institutional RAG setting. Unlike evaluations that focus primarily on retrieval metrics, the study jointly examines retrieval, generation and table-level structural preservation.

III. Experimental Setup

A. Dataset

The dataset consists of 65 publicly available institutional documents from Universiti Kebangsaan Malaysia (UKM). These documents include academic brochures, fee structures, program guidelines, and administrative policies. The corpus contains heterogeneous formats, overlapping terminology, and structured content such as tables and program codes. This makes it a practical testbed for evaluating chunking strategies in an institutional RAG setting. Table I summarizes the total documents included in the dataset based on categories:

TABLE I. Dataset Statistics

Category	No. of Documents
Students Undergraduate and Postgraduate Guidelines, Handbooks, Programme Brochures	21
Grant and Research Fund Guidelines	18
UKM ICT Guidelines	12
UKM's campus magazines	5
Open-source platform (Official UKM website)	2
UKM's town hall forum slides	3
Other UKM's centers guidelines	4
Total Documents: 65	

B. Hardware Setup

The experiments were conducted on a GMKTec AI PC equipped with the following specifications:

- GPU: AMD Radeon 8060s
- CPU: AMD Ryzen™ AI Max+ 395 (16 cores, up to 5.1GHz)
- RAM: 128GB
- Storage: 2TB

C. Evaluation Parameters

The evaluation is designed to assess system performance from three perspectives, including retrieval effectiveness, answer quality and structural integrity. The retrieval and generation settings are held constant across chunking configurations.

- Retrieval Evaluation: Measures the ability of the system to retrieve relevant document chunks.
- Answer Quality Evaluation: Assesses the correctness and relevance of generated responses.
- Structure-Sensitive Evaluation: Evaluates how well chunking strategies preserve structured information such as tables and program-specific data.

D. Query Dataset

A set of 28 evaluation queries was constructed to reflect realistic student information needs commonly encountered in institutional chatbot interactions. The queries are categorized into four types based on their information-seeking characteristics and structural complexity.

- Structured Queries: Questions requiring retrieval of tabular information, such as course fee amounts, program codes and semester durations. These queries test the system's ability to preserve co-located structured data across chunking strategies.
- Factual Queries: Questions seeking specific information from policy documents, guidelines, or institutional profiles. These queries evaluate retrieval precision and answer accuracy for well-defined information needs.
- Procedural Queries: Questions about multi-step processes, for example, application requirements or administrative procedures. These queries test the system's ability to retrieve and synthesize information spanning multiple document sections.
- Open-Ended Queries: Questions requiring comprehensive answers aggregating information from multiple sources.

Each query is paired with a ground truth answer derived from authoritative institutional documents, including fee structure tables, student handbooks, ICT guidelines, research funding policies and university policies. The query set comprises direct questions and paraphrased variants. All queries are written in Bahasa Melayu, reflecting the primary language of institutional documents and student interactions at

UKM. Ground truth answers range from short factual responses to paragraph-length explanations, with an average length of 87 words. The small number of queries, particularly the single open-ended query, limits the extent to which the results can support broad claims about diverse institutional question types. Table II summarizes the query dataset distribution by type and the percentage of each type.

TABLE II. QUERY DATASET STATISTICS

Query Type	Count	Percentage (%)
Structured	6	21.4
Factual	16	57.1
Procedural	5	17.9
Open-Ended	1	3.6
Total	28	100

E. Evaluation Metrics

The performance of the proposed system is evaluated using three types of metrics:

1) Retrieval metrics: Retrieval effectiveness is assessed using Hit@1, Hit@3, Precision@3, Recall@3, Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (nDCG). Hit@k is 1 when at least one relevant chunk appears among the top k retrieved chunks and 0 otherwise. Precision@3 is the proportion of the three retrieved chunks judged relevant, whereas Recall@3 is the proportion of all relevant chunks retrieved in the top three. MRR is the mean reciprocal rank of the first relevant chunk. nDCG measures ranking quality relative to the ideal ranking. The evaluation uses a fixed top-3 retrieval set for rank-based retrieval metrics, and after reranking, the top two chunks are then supplied to the generation model. These are the standard metrics for evaluating RAG retrieval performance as suggested in previous studies [7], [14], [17].

2) Answer quality metrics: Generated answers are evaluated using ROUGE-L, BLEU, BERTScore F1 and semantic similarity. ROUGE-L and BLEU measure lexical overlap with the ground-truth answer, while BERTScore and semantic similarity capture semantic alignment [11], [12], [19].

3) Structural integrity metric: To assess the preservation of structured information, a structural integrity metric is introduced. This addresses a gap in existing RAG evaluation frameworks [19], which do not explicitly measure table row co-location. This is a critical requirement for institutional documents that contain important information such as fee structures and program codes [9], [16].

IV. METHODOLOGY

This study implements a Retrieval-Augmented Generation (RAG) framework to investigate the impact of chunking strategies on retrieval and answer quality for structured institutional documents. The overall system architecture consists of three main stages: document preprocessing and chunking, dense retrieval and answer generation. The institutional documents in this study refer to the documents of Universiti Kebangsaan Malaysia (UKM) that are publicly available on the Internet.

A. Data Preprocessing Pipeline

Fig. 1 illustrates the data preprocessing pipeline involved in this study. Raw institutional PDFs are cleaned to remove irregular spacing, redundant headers and footers and unwanted characters. The cleaned content is then converted to text and segmented according to the evaluated chunking configuration. Each chunk is embedded using the intfloat/multilingual-e5-base model. This embedding model is suitable to be used for information retrieval or semantic similarity across various languages [20], which makes it a suitable model for this study, as most of the institutional documents are in the Malay language. This will convert the chunks into dense vector representations and stored in a local vector database. Finally, each text segment is enriched with metadata, including document source, document title and section information. This metadata provides contextual information that supports more effective retrieval and analysis.

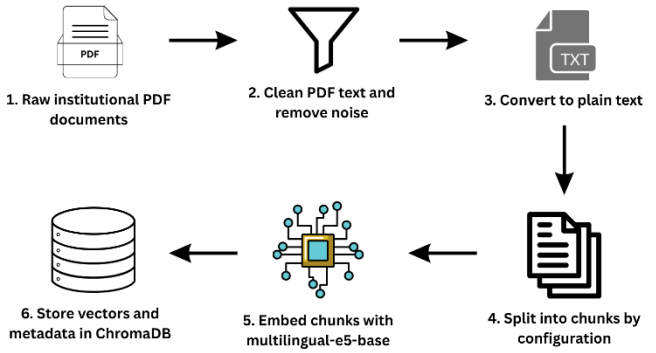

Fig. 1. Data preprocessing pipeline.

B. Retrieval-Augmented Generation Process

During the inference phase, a user query is embedded using the same embedding model used for document indexing. A similarity search retrieves the top three candidate chunks from the vector database. The BAAI/bge-reranker-v2-m3 model then reranks these candidates, and the top two chunks are supplied to the LLM together with the original query to generate the final response based on the retrieved knowledge and the user query.

To ensure experimental consistency and validity, the retrieval mechanism, embedding model, vector database configuration, reranker, language model and prompt template are kept constant across all experiments. Therefore, observed differences in retrieval accuracy or answer quality are primarily attributable to the chunking configuration.

C. Vector Store Setup and Retrieval Configuration

All chunks are embedded using the intfloat/multilingual-e5-base model, producing 768-dimensional dense vectors stored in ChromaDB, which is the chosen local vector database in this study. During query processing, the user query is embedded into the same vector space. Then, the top three (k=3) most similar chunks are retrieved and reranked using BAAI/bge-reranker-v2-

m3. The final top two (k=2) reranked chunks are then passed to the generation model for answer generation.

D. Generation Model

The retrieved chunks are provided as context to a large language model for answer generation. This study employs the Gemma-3-4b-it quantized model running locally via llama.cpp. The model was selected for its strong multilingual performance across more than 140 languages [21]. The generation parameters are as follows:

- Temperature: 0.5
- Max tokens: 512
- Top-p: 0.9

A single response was generated for each query under the fixed generation settings of the parameters. The same generation configuration was applied across all chunking configurations to maintain consistency. Because the model operates with non-zero temperature, the generation process may exhibit stochastic variation. However, repeated generations were not performed in the present experiment. Therefore, the reported answer-quality metrics represent performance from a single generation for each query and do not quantify generation-level variability.

E. Prompting Strategy

A consistent prompt template is used across all experiments to ensure fair comparison. Fig. 2 shows the prompt template used for generation. This detailed prompt instruction ensures the model relies solely on retrieved context and does not add external knowledge to reduce hallucinations.

```
"""Anda adalah pembantu AI yang pakar dalam Bahasa Melayu untuk menjawab
soalan tentang Universiti Kebangsaan Malaysia (UKM).
PERATURAN PENTING:
1. Jawab HANYA berdasarkan dokumen yang diberikan
2. Jika maklumat tidak ada dalam dokumen, jawab: "Maaf, tiada maklumat
dijumpai oleh Watan AI."
3. Berikan jawapan dalam SATU perenggan sahaja
4. Gunakan bahasa yang jelas dan profesional
5. JANGAN tambah maklumat dari pengetahuan luar
6. JANGAN ulang soalan atau label seperti "Jawapan:" atau "DOKUMEN:"
7. Berikan jawapan secara langsung

{context}

SOALAN: {query}

JAWAPAN:"""
```

Fig. 2. Prompt template.

F. Chunking Configurations

Four chunking strategies are evaluated across six configurations. The fixed-size strategy contains 256-token, 512-token and 512-token with 50-token overlap configurations, while the remaining strategies are recursive, semantic and structure-aware. Each configuration is used to construct a separate vector store while all other retrieval and generation settings remain unchanged. The fixed-size configurations represent selected small and moderate chunk sizes with an additional overlap condition rather than an exhaustive sensitivity analysis across chunk-size and overlap settings [15].

1) Fixed-size chunking: Fixed-size chunking divides text into uniform chunks based on a predefined token length. While computationally efficient, it does not consider semantic or structural boundaries, which may result in fragmented information [7], [12].

2) Recursive chunking: Recursive chunking splits text hierarchically based on natural linguistic boundaries such as paragraphs and sentences [7], [12]. This approach aims to maintain contextual coherence while controlling chunk size.

3) Semantic chunking: Semantic chunking groups sentences based on embedding similarity and meaning [12], ensuring that each chunk represents a semantically coherent unit of information [11],[13]. It can preserve topical coherence but may produce larger and more variable chunk sizes.

4) Structure-aware chunking: Structure-aware chunking is specifically designed for structured documents. It preserves meaningful structural elements such as tables, headings and program-specific data, ensuring that logically related information remains within the same chunk [16], [17]. This is particularly important for institutional documents containing important structural information such as fee tables and coded program information.

V. Results

A. Chunk Intrinsic Analysis

Before examining the retrieval and answer, the intrinsic properties of the chunking strategies were analyzed to understand their structural characteristics. Table III summarizes the total number of chunks produced, average chunk length in characters, intra-chunk semantic coherence and table integrity rate. Coherence is computed as the mean pairwise cosine similarity between sentence embeddings within a chunk, following established practice in semantic segmentation evaluation [13].

TABLE III. Chunk Intrinsic Statistics Across Chunking Configurations

Chunking Configuration	Total Chunks	Avg. Length	Coherence	Table Integrity
Fixed-size 256 tokens	10,011	177	0.813	0.469
Fixed-size 512 tokens	4,891	363	0.824	0.552
Fixed-size 512 tokens with 50-token overlap	4,925	368	0.824	0.523
Recursive	5,040	367	0.825	0.540
Semantic	1,517	1,172	0.816	**0.709**
Structure-aware	2,879	627	**0.826**	0.535

Fixed-size chunking at 256 tokens produced the largest number of chunks, which is 10,011 chunks with an average length of only 177 characters. This aggressive fragmentation results in the lowest intra-chunk coherence at 0.813 and the worst table integrity rate at 0.469, confirming that uniform, boundary-agnostic segmentation is poorly suited for documents containing structured tabular content [17]. Fixed-size chunking at 512 tokens reduced the chunk count to 4,891, improving average chunk length to 363 characters, coherence to 0.824, and table integrity to 0.552. Adding a 50-token overlap did not improve coherence but reduced table integrity marginally to

0.523, suggesting that overlapping boundaries may occasionally duplicate partial table rows across adjacent chunks.

Recursive chunking produced 5,040 chunks with an average length of 367 characters and the second-highest intra-chunk coherence at 0.825, consistent with the expectation that hierarchical splitting along natural linguistic boundaries preserves topical unity within each chunk [7]. Semantic chunking produced the fewest chunks, which are 1,517 and the largest average length (1,172 characters). This strategy achieves the highest table integrity rate at 0.709 because large chunks naturally encompass entire table sections. However, the lower coherence score of 0.816 relative to recursive and structure-aware strategies suggests that very large chunks may aggregate topically distinct content [13].

Structure-aware produced 2,879 chunks with an average length of 627 characters. This results in an intermediate between semantic and fixed-size approaches. The highest intra-chunk coherence overall, at 0.826 achieved through this chunking strategy. Its table integrity rate of 0.535 falls below that of semantic chunking, indicating that the heuristic-based table detection, while effective for well-formatted fee tables, does not consistently identify all table boundaries, particularly in documents using non-standard PDF layouts [9].

B. Retrieval Performance

Table IV presents the retrieval performance across six chunking configurations. The reported metrics include Hit@1, Hit@3, MRR, nDCG, Precision@3, and Recall@3.

TABLE IV. RETRIEVAL PERFORMANCE ACROSS CHUNKING CONFIGURATIONS

Chunking Configuration	Hit@1	Hit@3	MRR	nDCG	P@3	R@3
Fixed-size 256 tokens	0.893	0.893	0.893	0.893	0.559	0.297
Fixed-size 512 tokens	0.893	0.929	0.911	0.915	0.595	0.357
Fixed-size 512 tokens with 50-token overlap	**0.964**	**0.964**	**0.964**	**0.964**	**0.631**	**0.393**
Recursive	**0.964**	**0.964**	**0.964**	**0.964**	0.619	0.381
Semantic	0.929	**0.964**	0.946	0.951	**0.631**	**0.393**
Structure-Aware	0.893	**0.964**	0.929	0.938	0.619	0.381

C. Answer Quality

Table V presents answer quality results across ROUGE-L, BLEU, BERTScore F1 and semantic similarity. Unlike retrieval performance, substantial differences are observed across chunking configurations.

TABLE V. ANSWER QUALITY METRICS ACROSS CHUNKING CONFIGURATIONS

Chunking Configuration	ROUGE-L	BLEU	BERT Score	Semantic Sim.
Fixed-size 256 tokens	0.312	0.112	0.735	0.862
Fixed-size 512 tokens	0.464	0.227	0.786	**0.906**
Fixed-size 512 tokens with 50-token overlap	0.466	0.212	0.769	**0.906**
Recursive	0.456	0.199	0.764	0.903
Semantic	0.441	0.241	0.771	0.891
Structure-Aware	**0.492**	**0.263**	**0.788**	**0.906**

Structure-aware achieves the highest scores for ROUGE-L with a value of 0.492, BLEU with a value of 0.263 and BERTScore F1 with a value of 0.788. For the semantic similarity metric, the structure-aware chunking ties with both configurations of fixed-size chunking at 512 tokens, with overlap and no overlap, at the value of 0.906. These results indicate that answers generated from structure-aware chunks are both lexically closer and semantically equivalent to the ground truth, supporting the hypothesis that preserving structural relationships improves generation quality [16], [17].

Fixed-size chunking at 512 tokens ranks second overall. Adding a 50-token overlap slightly improves ROUGE-L to 0.466 but reduces the BLEU score to 0.212, suggesting that redundant content introduced by overlapping boundaries may reduce n-gram precision. Recursive chunking produces comparable results with ROUGE-L at the value of 0.456, while the BLEU score is 0.199. This chunking strategy performs marginally below structure-aware and fixed-size at the 512-token configuration. Meanwhile, semantic chunking yields a lower ROUGE-L value of 0.441 and a BERTScore value of 0.771 despite its strong retrieval characteristics. This situation occurs potentially because oversized chunks introduce irrelevant content that dilutes the language model's context window. This finding aligns with the observations by [11] that contextual completeness does not scale linearly with chunk size.

Fixed-size chunking at 256 tokens performs the worst across all answer quality metrics. This confirms that excessively small chunks are insufficient to provide the language model with adequate context for accurate answer generation [12], [15]. A critical observation from this study is the dissociation between retrieval and answer quality, where the fixed-size chunking at 256 tokens achieves the same Hit@1 (0.893) as fixed-512 and structure-aware strategies yet produces substantially inferior answer quality. This decoupling has been reported in prior RAG evaluations [12], [19] underscores the importance of evaluating RAG systems beyond retrieval metrics alone.

D. Structural Integrity Analysis

Table VI presents the structural-integrity rate, which measures whether fee amounts and programme codes are co-located with their required contextual fields (e.g., semester label, program name) within the same chunk. The metric is intended to address a limitation of standard RAG evaluation frameworks, which do not capture structural preservation of structured data relationships [19].

TABLE VI. TABLE INTEGRITY METRICS ACROSS CHUNKING CONFIGURATIONS

Chunking Configuration	Table Integrity
Fixed-size 256 tokens	0.469
Fixed-size 512 tokens	0.552
Fixed-size 512 tokens with overlap	0.523
Recursive	0.540
Semantic	**0.709**
Structure-aware	0.535

Semantic chunking achieves the highest table-integrity rate (0.709), while structure-aware achieves 0.535. This result reveals a limitation of the current structure-aware

implementation: the heuristic-based table detection relies on regular expression patterns that may not generalize to all table formats present in the heterogeneous UKM corpus. The fixed-size 256-token configuration exhibits the lowest table integrity (0.469), confirming that uniform segmentation at fine granularities is poorly aligned with the structural requirements of institutional documents [17], a finding consistent with structure-aware chunking evaluations in enterprise and scientific document settings [14], [15].

VI. Discussion

A. Retrieval Saturation and Its Implications

A consistent finding across all experiments is the convergence of retrieval metrics, with most strategies achieving Hit@3 of 0.964 and MRR values between 0.893 and 0.964. This saturation is likely attributable to the lexical specificity of institutional queries, which contain distinctive tokens such as program codes and fee amounts that facilitate accurate embedding-based retrieval regardless of chunk boundary decisions. Similar retrieval saturation has been reported in domain-specific RAG evaluations in clinical [12] and policy document settings [22] where query-document lexical overlap enables consistent top-k retrieval across different chunking configurations.

This finding challenges the common assumption that improving retrieval metrics is sufficient to improve downstream answer quality [7], [19]. The evidence presented demonstrates that two strategies may achieve identical Hit@1 scores while producing substantially different answer quality, as illustrated by the parity between fixed-256 and fixed-512 in Hit@1 (both 0.893) but divergence in ROUGE-L (0.312 vs. 0.464). This decoupling underscores the need for multi-dimensional evaluation frameworks, such as RAGAS [7], that assess faithfulness and answer correctness independently of retrieval rank. This study investigated the impact of different chunking strategies on retrieval and answer quality in RAG systems for structured institutional data.

B. The Case for Structure-Aware Chunking

Structure-aware chunking demonstrates consistent superiority in answer quality despite not leading in retrieval performance or structural-integrity scores. Its ROUGE-L score of 0.492 and BLEU score of 0.263 represent improvements of approximately 5.6% and 15.9%, respectively, over the next best strategy. These gains are attributable to the preservation of logical relationships within structured content, which provides the language model with coherent and complete contextual information [14], [17]. The intermediate chunk size (average 627 characters) strikes a balance between contextual richness for answer generation and retrieval specificity for similarity matching.

This finding correlates with prior work on structure-aware RAG in specialized domains. Taiwo et al. [17] reported that structure-aware chunking yielded higher retrieval effectiveness for enterprise oil and gas documents, particularly on top-k metrics, compared to fixed-size and semantic baselines. Similarly, a study on financial report chunking [14] demonstrated that element-type-based chunking, which preserves the semantic role of document components, improved RAG question answering accuracy compared to paragraph-level approaches. The present study extends these findings to multilingual institutional documents in Bahasa Malaysia, demonstrating that structure-aware chunking benefits generalize beyond English-language corpora. However, the evidence should not be interpreted as demonstrating universal superiority because the experiments use a single institutional corpus and a small query set.

C. Limitations of the Structure-Aware Approach

Despite its advantages in answer quality, the structure-aware chunker exhibits a lower table integrity rate (0.535) than semantic chunking (0.709). The table integrity gap reflects the inherent difficulty of detecting table boundaries in heterogeneous PDF-extracted text, where formatting cues such as tab characters or pipe delimiters are inconsistently preserved across document types. The current heuristic-based approach, while effective for well-formatted fee tables, may not generalize to all institutional document types present in the UKM corpus, such as campus magazine articles and policy guidelines with embedded tables in non-standard layouts. These limitations are consistent with observations in the broader structure-aware chunking literature. Amiri & Bocklitz [15] noted that domain-specific chunking strategies require careful calibration of boundary detection heuristics to avoid over-segmentation or under-segmentation. Recent work on topology-aware chunking [18] has proposed preserving hierarchical document structure through structured intermediate representations, suggesting a potential direction for improving the current heuristic implementation.

D. Chunk Size and Coherence Trade-offs

The intra-chunk coherence scores are closely clustered across all strategies (0.813 to 0.826), suggesting that coherence alone is insufficient to distinguish the configurations. However, when interpreted alongside chunk size, a clearer pattern emerges. The relationship between chunk size and answer quality is non-linear: answer quality improves from fixed-256 to fixed-512 and peaks at structure-aware but declines for semantic chunking despite its larger chunks. This non-linearity implies that an optimal chunk size range exists for this document collection, approximately 300 to 700 characters, beyond which the inclusion of extraneous content degrades generation quality. This observation is consistent with findings by researchers who showed that larger chunk sizes produced diminishing returns and diluted chunk coherence [11], and with reports of size-dependent performance trade-offs in chemistry and biomedical RAG settings [15].

E. Practical Implications for Institutional RAG Deployment

The findings carry practical implications for the design of RAG-based institutional chatbot systems. First, the consistent underperformance of fixed-size 256-token chunking across all answer quality and structural integrity metrics indicates that this approach should be avoided for document collections containing structured content, despite its computational efficiency [12], [17]. Second, structure-aware chunking is recommended for deployments where answer accuracy is prioritized, particularly in institutional contexts where erroneous fee or program information carries significant consequences for end users [23], [24]. Third, for resource-constrained deployments, fixed-512

chunking configuration without token overlapping represents a practical compromise, achieving the second-highest answer quality. Semantic chunking may be preferable when preservation of table-level structural integrity is the primary objective.

Several limitations constrain interpretation of the results. The evaluation uses 28 queries, including only one open-ended query. This situation may cause the results to be sensitive to the selected questions and should not be generalized broadly across institutional question types. The six configurations also do not constitute an exhaustive chunk-size sensitivity analysis. In addition, all documents are from UKM; thus, the application of the same data preparation pipeline for other institutional domains requires validation because document structures and terminology may differ. Finally, the generation settings use fix temperature of 0.5 and top-p 0.9, which permit stochastic output. Because repeated-generation variance is not reported, the presented generation metrics should be interpreted as point estimates rather than estimates of stability across repeated runs.

VII. Conclusion and Future Work

This study evaluated four chunking strategies across six configurations for RAG structured institutional documents. The results show that although retrieval performance tends to saturate across strategies, answer quality and structural integrity vary significantly. Structure-aware chunking demonstrates the strongest performance in answer quality, while semantic chunking achieves the highest table-integrity score. These findings highlight that no single chunking configuration dominates every evaluation objective and that chunking should be selected according to the target document structure and application objective. Future work should expand the query set, conduct systematic chunk-size sensitivity analysis, evaluate repeated generations to quantify stochastic variability and test the approaches on other institutional documents.

Acknowledgment

The authors want to thank the Ministry of Higher Education (MoHE) Malaysia and Universiti Kebangsaan Malaysia under Grant Code: WARISAN-2025-001.

References

[1] B., Sai & U., Vyshnavi & Y., Kavya & R., Sai & G., Ramya, Intelligent FAQ Chatbot: A User-Centric Approach using Large Language Models. Journal of Artificial Intelligence and Capsule Networks. 7. 78-93, 2025.

[2] Husain, M. L., Wibisono, Y., & Anisyah, A. Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAG). Brilliance: Research of Artificial Intelligence, 5(2), 727–735, 2025.

[3] Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology, 19(5), 598–604, 2023.

[4] Santoso, H. A., Saraswati, G. W., Rohman, M. S., Anisa, N., Winarsih, S., Sukmana, S. E., Nugraha, A., Mulyanto, E., Rustad, S., & Firdausillah, F. Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services. International Seminar on Application for Technology of Information and Communication (ISemantic), 417–423. 2018.

[5] Swacha, J., & Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. Applied Sciences (Switzerland), 15(8), 2025.

[6] Adamopoulou, E., & Moussiades, L. An Overview of Chatbot Technology. IFIP Advances in Information and Communication Technology, 584 IFIP, 373–383, 2020

[7] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. http://arxiv.org/abs/2312.10997, 2024.

[8] Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. IEEE Transactions on Education, 68(1), 103–116, 2025.

[9] Zhang, Q., Wang, B., Huang, V. S.-J., Zhang, J., Wang, Z., Liang, H., He, C., & Zhang, W. Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. http://arxiv.org/abs/2410.2116, 2024

[10] Hui, Y., Lu, Y., & Zhang, H. UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis. Advances in Neural Information Processing Systems, 67200–67217, 2024.

[11] Merola, C., & Singh, J. Reconstructing Context: Evaluating Advanced Chunking Strategies for Retrieval-Augmented Generation. http://arxiv.org/abs/2504.19754, 2025

[12] Gomez-Cabello, C. A., Prabha, S., Haider, S. A., Genovese, A., Collaco, B. G., Wood, N. G., Bagaria, S., & Forte, A. J.. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. Bioengineering, 12(11), 2025.

[13] Kiss, C., Nagy, M., & Szilágyi, P. Max–Min semantic chunking of documents for RAG application. Discover Computing, 28(1), 2025.

[14] Cheerla, C.. Advancing Retrieval-Augmented Generation for Structured Enterprise and Internal Data. http://arxiv.org/abs/2507.12425. 2025

[15] Amiri, M., & Bocklitz, T. Chunk Twice, Embed Once: A Systematic Study of Segmentation and Representation Trade-offs in Chemistry-Aware Retrieval-Augmented Generation. http://arxiv.org/abs/2506.17277, 2025.

[16] Yu, X., Jian, P., & Chen, C. TableRAG: A Retrieval Augmented Generation Framework for Heterogeneous Document Reasoning. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 14063–14082, 2025.

[17] Taiwo, S., & Amaluddin Yusoff, M. Evaluating Chunking Strategies For Retrieval-Augmented Generation In Oil And Gas Enterprise Network Security&Applications, 49–67, 2026.

[18] Liu, X. TopoChunker: Topology-Aware Agentic Document Chunking Framework. http://arxiv.org/abs/2603.18409, 2026.

[19] Es, S., James, J., Espinosa-Anke, L., Schockaert, S., & Gradients, E. RAGAS: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158, 2024.

[20] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F.. Multilingual E5 Text Embeddings: A Technical Report. http://arxiv.org/abs/2402.05672, 2024.

[21] Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. Gemini: A Family of Highly Capable Multimodal Models. http://arxiv.org/abs/2312.11805, 2025.

[22] Latif, S., Ameer, H., Akram, M. H., & Fatima, M. The Chunking Paradigm: Recursive Semantic for RAG Optimization. Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025), 137–145, 2025.

[23] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. Survey of Hallucination in Natural Language Generation. In ACM Computing Surveys (Vol. 55, Number 12). Association for Computing Machinery, 2023.

[24] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 9459–9474, 2020.