MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 11 | No. 1 | Halaman 88 - 101
DOI : https://doi.org/10.26760/mindjournal.v11i1.88-101 Juni 2026

The Impact of Chunking Granularity on Hybrid GraphRAG Architecture Performance in Mitigating Hallucinations

AFIN MAULANA, YUSUP MIFTAHUDDIN*, DIASH FIRDAUS

[1,]Department of Informatics, Institut Teknologi Nasional Bandung, Indonesia
Email: yusufm@itenas.ac.id

Received 14 April 2026 **|** *Revised* 14 Mei 2026 **|** *Accepted* 15 Juni 2026

ABSTRAK

Pesatnya pertumbuhan literatur herbal memicu information overload yang menghambat ekstraksi data manual. Meskipun Large Language Models (LLMs) membantu otomasi, risiko halusinasi faktual pada domain medis tetap tinggi, sementara Retrieval-Augmented Generation (RAG) konvensional sering gagal menangkap hubungan relasional antar-entitas. Penelitian ini menerapkan Hybrid GraphRAG, menggabungkan pencarian vektor dan Knowledge Graph, untuk mengatasi kelemahan tersebut. Fokus utamanya adalah menguji dampak granularitas chunking (karakter, kata, kalimat) terhadap representasi pengetahuan, mengingat fragmentasi teks berisiko memutus konteks semantik. Hasil eksperimen menunjukkan bahwa chunking berbasis kalimat memberikan performa terbaik, menggandakan skor Correctness dan Recall dari 0,28 ke 0,56. Temuan ini menegaskan pentingnya menjaga keutuhan kalimat demi akurasi dan keterhubungan data dalam sistem informasi medis.

***Kata kunci**: Hybrid GraphRAG, Knowledge Graph, Chunking, Tanaman Herbal*

ABSTRACT

The rapid growth of herbal medicine literature triggers an information overload that hinders manual data extraction. Although Large Language Models (LLMs) assist in automation, the risk of factual hallucination within the medical domain remains high, while conventional Retrieval-Augmented Generation (RAG) frequently fails to capture relational connections between entities. To address these limitations, this study implements a Hybrid GraphRAG architecture that integrates vector search and Knowledge Graphs. The primary focus is to evaluate the impact of chunking granularity (character, word, and sentence-level) on knowledge representation, considering that text fragmentation risks disrupting semantic context. Experimental results demonstrate that sentence-based chunking yields the best performance, doubling the Correctness and Recall scores from 0.28 to 0.56. These findings emphasize the importance of preserving sentence integrity for data accuracy and interconnectivity within medical information systems.

***Keywords**:Hybrid GraphRAG, Knowledge Graph, Chunking, Herbal Plants*

1. INTRODUCTION

The rapid expansion of unstructured text data presents a significant challenge in data science and Natural Language Processing, specifically in transforming raw data into structured knowledge **(Zubiaga, 2023)**. While efficient information extraction is vital across various sectors, it is especially critical in the herbal domain, where scientific literature has grown exponentially **(Salmerón-Manzano et al., 2020)**. This field involves complex, cross-disciplinary links between plants, compounds, and pharmacological effects that make manual synthesis impractical. Although Large Language Models (LLMs) offer advanced context understanding, their tendency toward factual hallucinations poses substantial risks in medical contexts **(Ji et al., 2024)**. Consequently, there is an urgent need for automated systems that can navigate these complexities without compromising accuracy.

In response to these hallucinations, Retrieval Augmented Generation (RAG) architecture was developed to combine the generative capabilities of LLMs with document retrieval from trusted external sources **(Lewis et al., 2021)**. This approach enables models to produce context-based answers that improve information reliability. Nevertheless, conventional RAG still relies on vector similarity that processes facts in isolation, failing to capture explicit relationships between entities.

Research by **Firdaus et al., 2024** indicates that conventional RAG in the Indonesian herbal domain possesses both potential and limitations. While it improves relevance compared to LLMs without retrieval, performance remains inconsistent across different models. Mistral 7B achieved higher METEOR scores than LLaMA2 7B but showed lower ROUGE precision. These findings suggest that improved text cohesion does not always guarantee factual accuracy. Furthermore, conventional RAG struggles to maintain stable relational connections within the herbal domain, which involves complex links between plants, active compounds, and pharmacological effects.

To overcome the limitations of conventional RAG, hybrid approaches integrating knowledge graphs with vector based retrieval have emerged. One implementation is HybridRAG, which combines VectorRAG and GraphRAG to enhance information extraction from unstructured data **(Sarmah et al., 2024)**. This method demonstrates improved performance in generating relevant and contextual answers compared to using vector or graph methods separately, as it leverages the relational representation of knowledge graphs and the semantic search power of vectors. However, the effectiveness of this approach depends heavily on the quality of information generated during the initial processing stages.

This dependency makes chunking strategy a vital part of the hybrid RAG architecture. **Barnett et al., 2024** showed that these strategies directly affect retrieval performance in vector-based systems. In a hybrid setup, chunking is even more critical because the text segments are used for both semantic search and extracting relations for the knowledge graph. Using rigid limits like strict character or word counts can break natural semantic boundaries and lower retrieval quality **(Gao et al., 2024)**. This fragmentation often separates related information like a plant name and its medical benefits into different chunks. Such gaps cause a loss of relational context and create isolated entities that weaken the Knowledge Graph's integrity. While sentence-based chunking keeps the context intact, it causes variations in information density. Currently, there is no systematic study on how these text boundaries simultaneously affect both vector and graph representations within a Hybrid GraphRAG architecture.

Based on the issue of losing relational context due to text fragmentation, the proposed study aims to analyze the influence of character, word, and sentence based chunking granularity on

Hybrid GraphRAG system performance in the herbal domain. Evaluation utilizes ROUGE metrics (precision, recall, and F-measure) and METEOR to assess linguistic quality, alongside Ragas metrics such as Correctness to measure factual consistency. This research focuses on evaluating whether the chunking boundaries affect hallucination mitigation, vector representation accuracy, and relational connectivity. The results are expected to identify the most effective strategy and provide insights into how granularity variations influence the underlying mechanisms of the proposed system.

2. RESEARCH METHODOLOGY

This research methodology outlines the systematic steps taken to analyze the impact of chunking granularity on Hybrid GraphRAG system performance within the herbal domain. The stages include architectural design, dataset preparation, implementation of chunking strategies, construction of knowledge representations in vector and graph forms, context integration, and system performance evaluation procedures.

2.1 Hybrid GraphRAG Architecture

The Hybrid GraphRAG system architecture is designed to combine the advantages of vector based semantic search with relational navigation through a knowledge graph.

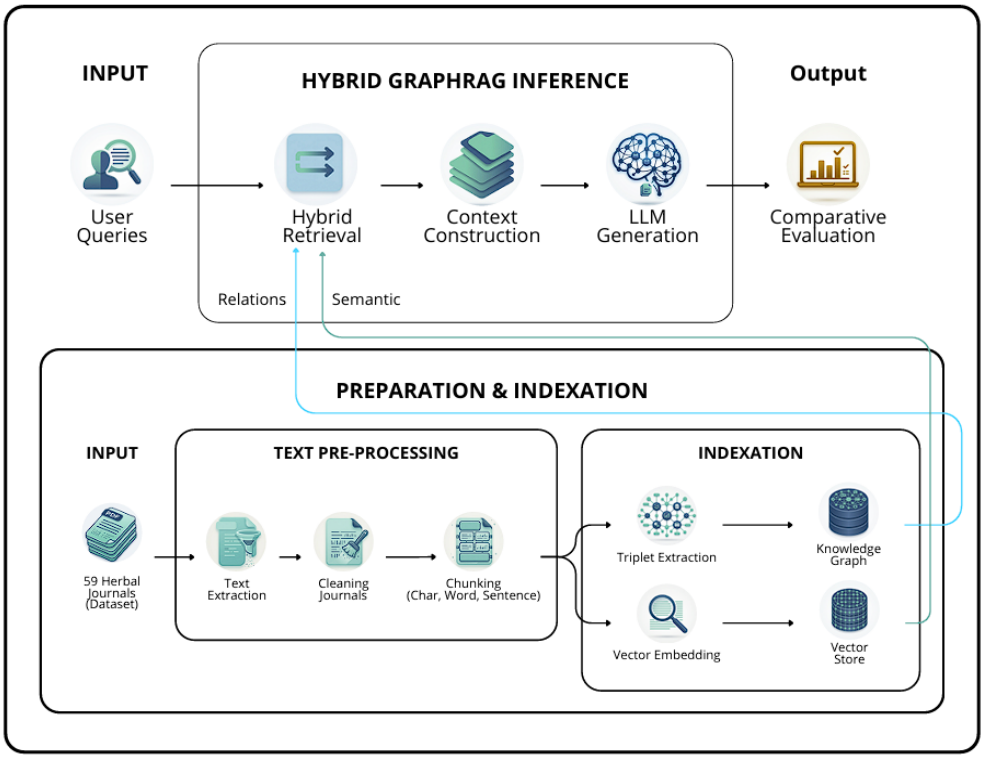

Figure 1. Hybrid GraphRAG System Architecture

The system workflow in Figure 1 begins with the preparation and indexing stage through the extraction of 50 herbal journals using PyMuPDF. Raw text is cleaned of noise elements such as metadata and headers using the Mistral 7B model and Regex normalization. The processed text is divided using character, word, and sentence-based chunking mechanisms to form knowledge representations in BGE-Small-v1.5 vector embeddings and a knowledge graph through triplet extraction.

During the inference stage, user queries are processed through hybrid retrieval that integrates vector semantic search and graph relation mapping. The resulting context serves as the basis for Mistral 7B to generate the final response. Finally, the system undergoes an output stage for performance evaluation using ROUGE, METEOR, and Ragas metrics to test the linguistic quality and factual consistency of the generated answers.

2.2 Data Description

This research utilizes a text corpus of Indonesian herbal plants consisting of 59 scientific journal documents in PDF format published between 2014 and 2025. The dataset is divided into two primary groups consisting of 9 reference journals adopted from the research by **Firdaus et al. (2024)** and 50 additional scientific articles from the Traditional Medicine Journal of Universitas Gadjah Mada. Details regarding the 9 primary knowledge base journals are presented in Table 1.

Table 1. List of Herbal Knowledge Base Journals

No	Author and Year	Journal Title
1	Sianipar (2021)	The Potential of Indonesian Traditional Herbal Medicine as Immunomodulatory Agents A Review
2	Putri et al. (2016)	Ethnobotanical study of herbal medicine in Ranggawulung Urban Forest Subang District West Java Indonesia
3	Fathir et al. (2021)	Ethnobotanical study of medicinal plants used for maintaining stamina in Madura ethnic East Java Indonesia
4	Sholikhah (2016)	Indonesian medicinal plants as sources of secondary metabolites for pharmaceutical industry
5	Ardiyanto et al. (2021)	The use of hyperuricemia herbs at Hortus Medicus herbal medicine clinic Tawangmangu
6	Elfahmi et al. (2014)	Jamu Indonesian traditional herbal medicine towards rational phytopharmacological use
7	Arozal et al. (2020)	Selected Indonesian Medicinal Plants for the Management of Metabolic Syndrome Molecular Basis and Recent Studies
8	Kartini et al. (2019)	Standardization of Some Indonesian Medicinal Plants Used in Scientific Jamu
9	Sumarni et al. (2019)	The scientification of jamu a study of Indonesians traditional medicine

The selection of the 9 reference journals was based on a rigorous selection process by herbal experts in previous research to ensure data relevance and suitability across various Indonesian herbal plants. Furthermore, the addition of 50 articles from Universitas Gadjah Mada was conducted to deepen the information coverage. This source was chosen due to its strong academic reputation and its presentation of consistent and systematically organized data structures. The integration of expert validated journals and structured additional literature aims to support the formation of complex relations within the knowledge graph and ensure the system possesses in depth knowledge coverage.

2.3 Chunking Strategies

The chunking strategy in this study serves as the primary variable significantly influencing the quality of knowledge representation within the Hybrid GraphRAG system. This process evaluates three approaches namely character based, word based, and sentence based methods to divide text into smaller units to support retrieval and in depth information extraction. Chunk formation is executed dynamically with a maximum limit of 512 tokens to align with the capacity of the embedding model used.

Table 2. Chunk Distribution Statistics

Approach	Total Chunk	Average Tokens	Min	Max	Total Token
Character	715	476,04	17	512	340,365
Word	705	490,82	4	497	346,030
Sentence	754	459,05	10	512	346,125

Based on Table 2, statistical analysis reveals unique characteristics of each strategy. The sentence based approach produces the highest number of chunks with 754 units and an average of 459,05 tokens. The character based strategy generates 715 chunks with an average of 476.04 tokens, while the word based strategy shows the highest average at 490.82 tokens with the fewest total chunks at 705 units. All three methods demonstrate a wide distribution range, with minimum values starting from 4 tokens up to the maximum limit of 512 tokens. These distributional differences reflect variations in information density, reinforcing the hypothesis that chunking granularity affects the effectiveness of vector based retrieval and the integrity of knowledge graph relations in distinct ways.

2.4 Knowledge Representation

In the Hybrid GraphRAG architecture, knowledge representation is constructed through two primary approaches namely vector based and knowledge graph based. Vector based representation is utilized to capture semantic similarity between text segments, while the knowledge graph is used to represent explicit relationships between entities. These two approaches complement each other in providing a more comprehensive context during the retrieval process, thereby enabling the system to generate more relevant and structured answers.

2.4.1 Vector Database

Vector based representation is performed by converting each text chunk into an embedding using the BGE Small v1.5 model. This model was selected due to its strong performance in semantic similarity tasks and its computational efficiency for processing large volumes of documents. Each embedding is represented within a 384 dimensional vector space.

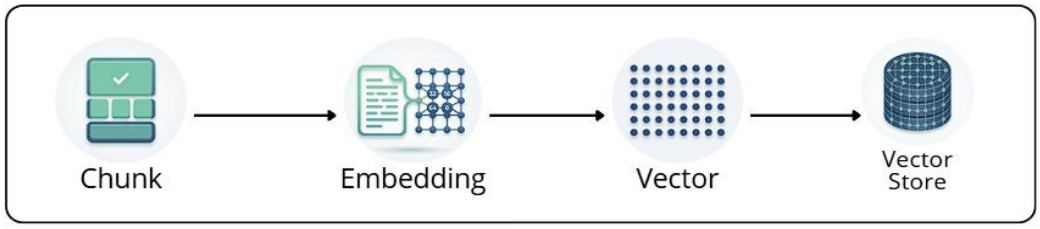

Figure 2. Vector Based Representation Workflow

Based on Figure 2, each chunk is processed by the embedding model to generate a vector representation which is subsequently stored in a vector store as a search index. During the retrieval stage, the user query is converted into an embedding using the same model, followed by a search based on cosine similarity to identify the most relevant chunks. This approach is effective in capturing implicit meaning similarities but remains unable to represent explicit relationships between entities, thus requiring integration with a graph based approach.

2.4.2 Knowledge Graph

Knowledge graph based representation is constructed to extract structured information from the herbal corpus through a 2-tier extraction chain approach using the Mistral 7B model. This approach adopts the methodology from **Sarmah et al. (2024)** which divides the extraction process into two stages to transform narrative text into relational facts in the form of triplets incrementally.

Figure 3. 2-Tier Extraction Workflow for KG Construction

As shown in Figure 3, the process begins with text chunks that are summarized into technical representations in the first stage, focusing on essential entities such as plants, active compounds, disease symptoms, and medical benefits. This stage aims to simplify the context and eliminate irrelevant information. Subsequently, in the second stage, the summarized results are processed to explicitly identify entities and relationships between entities, producing formal triplets organized into a knowledge graph with entities as nodes and relations as edges.

2.5 Hybrid GraphRAG Mechanism

The Hybrid GraphRAG mechanism integrates vector based and knowledge graph based representations through a dual path retrieval process that complements each other. This approach aims to combine the strengths of semantic search and relational navigation to build a comprehensive context before the answer generation process.

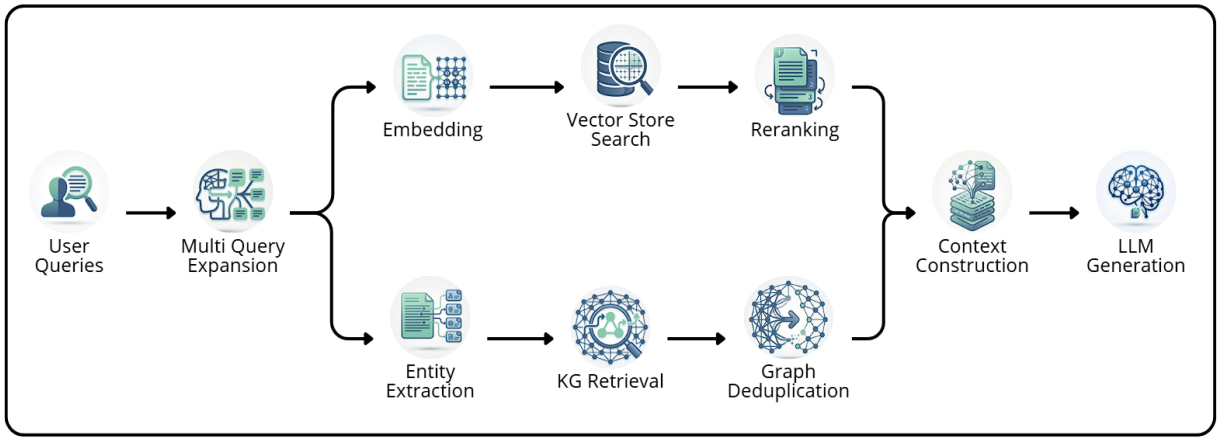

Figure 4. Hybrid GraphRAG Mechanism

Based on Figure 4, the user query is initially processed using the Multi Query Expansion (MQE) technique. This method expands a single primary query into multiple alternative variations to capture a wider range of potential keywords and semantic nuances, thereby maximizing the system's retrieval coverage. These expanded queries then enter two parallel search paths. In the vector search path, the queries are converted into embeddings using the BGE-Small-v1.5 model and undergo a rigorous Reranking process to ensure high precision, ultimately retrieving the five most relevant narrative text segments, or Top-5 chunks.

Simultaneously, the second path executes a knowledge graph traversal by extracting key entities to retrieve relevant relational triplets from the database. This set of triplets is subsequently processed through a Graph Deduplication stage to ensure data integrity. This deduplication process is essential for identifying and removing redundant or overlapping entity relations that may arise from extraction across various document fragments. The output of this stage is a refined collection of structured facts that are concise, unique, and entirely free from informational redundancy.

The retrieval results from both paths are integrated during the Context Construction phase. At this stage, the Top-5 narrative chunks are paired with the set of graph triplets and formatted together into a unified instruction template. This combined context is then provided to the Mistral 7B model to synthesize the final response. The selection of Mistral 7B is based on the findings of **Firdaus et al. (2024)**, which demonstrate its superiority in extracting facts within the Indonesian herbal domain compared to LLaMA2 7B. The integration of both textual and relational contexts ensures that the model receives comprehensive information, thereby effectively minimizing hallucinations.

2.6 Experimental Design

The experimental design in this study aims to evaluate how chunking granularity influences Hybrid GraphRAG system performance by comparing three primary methods, namely vector based, knowledge graph based, and hybrid retrieval. Each configuration is tested using three chunking strategies (character, word, and sentence) to analyze the impact of granularity on each knowledge representation mechanism.

Evaluation is conducted using 6 queries adopted from previous research, which have been validated alongside ground truth by experts to ensure a credible assessment basis. Each query represents various information needs within the herbal domain, with specific details of the queries and ground truth presented in Table 3.

Each method combination is tested 5 times to accommodate the generative nature of the language model, ensuring that evaluation results are not influenced by single output variations. The testing encompasses three primary scenarios namely vector based retrieval, knowledge graph, and Hybrid GraphRAG. Within the vector and knowledge graph scenarios, every chunking strategy is tested independently to observe the direct impact on the information extraction process. Meanwhile, in the hybrid scenario, chunking strategies are aligned across pathways to maintain contextual consistency during the integration of both representations. All testing results are subsequently averaged for each combination to obtain stable evaluation values.

Table 3. List of Questions and Ground Truth

No	Question	Ground Truth
Q1	Herb for headache	Cinnamon is a spice that has anti-inflammatory and neuroprotective properties. Researchers were therefore interested in studying whether cinnamon could help reduce migraine attacks and inflammation. For example, this journal describes about that
Q2	Herb For Diabetes	Trigonella foenum-graecum is one of the important medicinal plants in the management of diabetes mellitus. Several studies, such as (Geberemeskel et al., 2019), have investigated the effect of Trigonella foenum-graecum seed powder on the lipid profile of newly diagnosed type II diabetic patients.
Q3	What Herb for hypertension	Hypertension is a disease that is quite high in Indonesia, a major risk factor for cardiovascular disease (CVD). One of the herbs used is Centella asiatica (L) Urb. belongs to the Apiaceae (Umbelliferae) plant family, which has high triterpenoids and flavonoids and has antioxidant properties and is involved in the reninangiotensin-aldosterone system, which is an important hormonal system for blood pressure regulation.
Q4	Medical herb for fever	that A. manihot and its bioactive constituents have a wide range of biological properties, including anti-diabetic nephropathy, antioxidant, antiadipogenic, anti-inflammatory, analgesic, anticonvulsant, antidepressant, antiviral, antitumor, cardioprotective, antiplatelet, neuroprotective activity, immunomodulatory, and hepatoprotective. And A.manihot can be used for fever. However, further studies and clinical trials are still needed to confirm these findings.
Q5	Medical herb for rheumatism	Rheumatoid arthritis is one part of rheumatic disease. In Indonesia, some herbs that are often used for rheumatism are jambe jackfruit, and several other examples, such as cinnamon, curcumin, African tree, and so on.
Q6	*Medical herb for Heartburn*	Some sources have been researched, such as Harvard Medical School, which states that ginger root is a popular herbal remedy for heartburn. It has been used for centuries to relieve the symptoms of heartburn, such as a burning sensation in the chest

2.7 Evaluation Metrics

System performance evaluation is conducted using a combination of text similarity metrics and factual assessment to measure the impact of chunking granularity on the Hybrid GraphRAG architecture. The primary focus of this evaluation is to ensure that system answers are accurate both linguistically and factually within the Indonesian herbal domain.

Linguistic quality is measured through ROUGE N and METEOR metrics. Precision, recall, and F1-measure parameters in ROUGE are utilized to assess information accuracy and completeness based on n gram matching with established ground truth. Meanwhile, METEOR evaluates language quality more flexibly through synonym matching and stemming.

Factual accuracy aspects are assessed using the Answer Correctness metric from the Ragas framework. This metric utilizes a language model to measure the alignment of answer claims against original references to detect hallucinations. All metrics used have a value range from

0 to 1, where higher scores approaching 1 indicate better and more accurate system performance.

3. RESULTS AND DISCUSSION

This section presents the implementation results and evaluation of the Hybrid GraphRAG system within the Indonesian herbal domain through quantitative and qualitative analysis. The discussion begins with the characteristics of the knowledge base formed from various chunking strategies and continues with testing system performance in generating factual answers aligned with expert references. All data presented are results of controlled observations to identify the most optimal architectural configuration for handling herbal knowledge queries.

3.1 Knowledge Base Statistics

The knowledge base construction stage transforms the herbal corpus into vector and graph representations for each chunking strategy. Differences in text segmenting mechanisms influence the density of information indexed within the system. A summary of the database construction results is presented in Table 4 as evidence of the Hybrid GraphRAG architecture readiness before performance testing.

Table 4. Summary of Knowledge Base Statistics

Chunking Strategy	Total *Chunks*	KG Nodes	KG Edges
Character	715	2.952	4.641
Word	705	2.832	4.536
Sentence	754	3.052	4.741

The data presented in Table 4 indicates that each strategy generates varying volumes of data within the graph structure, with the sentence-based approach yielding the highest number of data units and entities (KG Nodes). This suggests that utilizing complete sentence units is more effective at preserving contextual integrity. Conversely, the reduction in data volume observed in the character and word strategies occurs because text fragmentation often disrupts the inherent meaning of a word or phrase. Consequently, the model loses critical context and fails to fully extract relationships between entities. Therefore, the sentence strategy proves to be the most effective in maintaining semantic quality and constructing comprehensive medical information relations.

The extraction results in Table 5 demonstrate the system ability to precisely identify entities and relations through the 2-tier extraction chain. The varied cutoff points across strategies significantly impact information completeness. While the character and word strategies extract basic facts despite stopping prematurely at the letters "ev" or the word "inhibit", only the sentence strategy successfully captures the complete context such as cancer cell inhibition. To better illustrate how this raw narrative text transforms into a graph representation for the retrieval pathway, Figure 5 provides a visual snapshot of the generated knowledge graph.

Table 5. 2-Tier Extraction Process in KG Construction

Strategy	Chunk (Raw)	Tier 1: Summary	Tier 2: Triplet
Sentence	*Curcuma longa, commonly known as turmeric, has been widely studied for its potential in treating gastrointestinal disorders, particularly due to its anti-inflammatory properties that may alleviate stomach discomfort and even inhibit cancer cell growth.*	Turmeric treats gastrointestinal disorders, has anti-inflammatory properties, and inhibits cancer cell growth	(Turmeric, treats, gastrointestinal disorders), (Turmeric, has, anti-inflammatory properties), (Turmeric, inhibits, cancer cell growth)
Word	Curcuma longa, commonly known as turmeric, has been widely studied for its potential in treating gastrointestinal disorders, particularly due to its anti-inflammatory properties that may alleviate stomach discomfort and even inhibit	Turmeric treats gastrointestinal disorders and has anti-inflammatory properties	(Turmeric, treats, gastrointestinal disorders), (Turmeric, has, anti-inflammatory properties)
Character	Curcuma longa, commonly known as turmeric, has been widely studied for its potential in treating gastrointestinal disorders, particularly due to its anti-inflammatory properties that may alleviate stomach discomfort and ev	Turmeric treats gastrointestinal disorders and has anti-inflammatory properties	(Turmeric, treats, gastrointestinal disorders), (Turmeric, has, anti-inflammatory properties)

Based on Figure 5, the medicinal plants *Euphorbia tirucalli* and *Cassia alata* are explicitly connected to their shared active compounds and specific disease conditions. This visualization demonstrates how the GraphRAG concept connects isolated facts into a structured relational network. By linking multiple plants through shared chemical properties, the system strengthens the semantic context and analytical capability for the subsequent hybrid retrieval process.

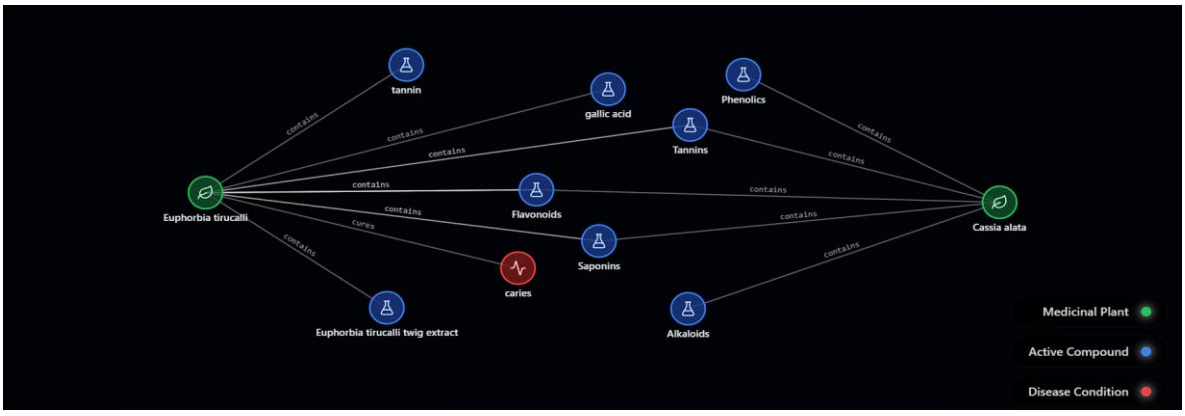

Figure 5. Subgraph Visualization

3.2 Evaluation Results

The quantitative evaluation stage is conducted to measure the effectiveness of the Hybrid GraphRAG architecture compared to pure vector based and pure graph based retrieval mechanisms. This testing uses average values from five iterations across six test queries to ensure system performance stability. Table 6 presents the recapitulation of evaluation scores based on ROUGE N, METEOR, and Answer Correctness metrics for every combination of chunking strategy and retrieval method.

Table 6. RAG System Performance Evaluation Results

Chunk	Retrieval Method	Precision	Recall	F-Measure	Meteor	Correctness
Character	Baseline	0,07	0,23	0,10	0,22	0,28
	Vector	0,15	0,39	0,21	0,23	0,42
	Graph	0,06	0,54	0,11	0,16	0,37
	Hybrid GraphRAG	0,15	0,50	0,22	0,26	0,51
Word	Vector	0,15	0,40	0,21	0,25	0,44
	Graph	0,06	0,54	0,11	0,16	0,31
	Hybrid GraphRAG	0,13	0,48	0,20	0,25	0,46
Sentence	Vector	0,16	0,41	0,22	0,26	0,50
	Graph	0,07	0,55	0,12	0,18	0,40
	Hybrid GraphRAG	**0,16**	**0,56**	**0,25**	**0,30**	**0,56**

Based on Table 6, Hybrid GraphRAG using the sentence based strategy achieves the highest performance across all metrics. An Answer Correctness score of 0.56 outperforms the Baseline which reaches 0.28. This evidence confirms that sentence units maintain optimal context for the integration of semantic search and graph relations. Conversely, word and character strategies reduce quality as information structures become fragmented. Nevertheless, all Hybrid GraphRAG variations continue to exceed Baseline scores, particularly in the recall aspect. The decline in scores at finer granularities confirms that the loss of contextual integrity hinders answer synthesis accuracy.

3.3 Analysis of Chunking Impact and Hybrid Retrieval

Evaluation results indicate that chunking granularity plays a fundamental role in system performance. The sentence based strategy consistently yields the best performance due to its ability to maintain semantic context in its entirety, thereby enhancing knowledge representation quality and facilitating the model in capturing relationships between entities. Conversely, word and character based methods tend to fragment context excessively, leading to limited relational structures within the knowledge graph and reduced retrieval effectiveness. These findings emphasize that chunking is not merely a technical optimization but a foundation that determines the quality of information for further processing.

Regarding retrieval, Hybrid GraphRAG proves to provide the most balanced performance by combining the strengths of vector based retrieval, which excels in precision and correctness, with graph based retrieval that offers high recall. This integration creates a complementary

combination where information accuracy and coverage can be optimized simultaneously. However, the effectiveness of this configuration depends heavily on the quality of the chunks used. Chunks that are too long potentially increase noise and exceed embedding token limits, preventing important information from being well represented. Consequently, optimal system performance can only be achieved through the synergy between appropriate chunk granularity and the hybrid retrieval mechanism.

4. CONCLUSION

This research demonstrates that the Hybrid GraphRAG method effectively improves the quality of question answering systems within the Indonesian herbal domain through the integration of semantic and relational representations. The best performance was achieved in the sentence based chunking scenario with a correctness value of 0.56 and recall of 0.56, reflecting the system ability to generate accurate answers while maintaining broad information coverage. These findings confirm that knowledge representation quality at the chunking stage serves as a key factor in supporting retrieval effectiveness.

Furthermore, the implementation of a 2 tier extraction chain plays a vital role in maintaining knowledge graph quality, particularly in ensuring the relevance and consistency of extracted entities and relations. The combination of vector and graph pathways produces an adaptive retrieval mechanism where the system balances information precision and completeness simultaneously. Thus, the hybrid configuration not only addresses the limitations of single methods but also enhances system performance consistency when handling queries requiring relational reasoning.

Future research should expand the evaluation by involving multidisciplinary experts, such as botanists and pharmacologists, to conduct a deeper cross-expert validation. This step will significantly improve the accuracy and reliability of the ground truth used for system evaluation. Furthermore, incorporating larger herbal datasets and testing the architecture with various other Large Language Models is highly recommended to strengthen system stability. If the system is to be adapted for different literature domains, specific adjustments will be necessary. These adjustments include modifying the prompt instructions for entity extraction and tuning the text chunking limits to align with the context window capacity of the chosen models, ensuring that the system can generalize effectively across a wider knowledge scale.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Institut Teknologi Nasional Bandung for providing the facilities and resources necessary for this research. The authors also acknowledge the financial and institutional support provided by the Institute for Research and Community Service (LPPM), Institut Teknologi Nasional Bandung.

REFERENCES

Ardiyanto, D., Triyono, A., Nisa, U., Fitriani, U., Astana, P. R., Novianto, F., & Zulkarnain, Z. (2021). The use of hyperuricemia herbs at "Hortus Medicus" herbal medicine clinic Tawangmangu. *JKKI : Jurnal Kedokteran Dan Kesehatan Indonesia*, *12*(2), 158–165. https://doi.org/10.20885/JKKI.Vol12.Iss2.art9

Arozal, W., Louisa, M., & Soetikno, V. (2020). Selected Indonesian medicinal plants for the management of metabolic syndrome: Molecular basis and recent studies. *Frontiers in cardiovascular medicine*, *7*, 82. https://doi.org/10.3389/fcvm.2020.00082

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024, April). Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* (pp. 194-199). https://doi.org/10.1145/3644815.3644945

Fathir, A., Haikal, M., & Wahyudi, D. (2021). Ethnobotanical study of medicinal plants used for maintaining staminain Madura ethnic, East Java, Indonesia. *Biodiversitas*, *22*(1), 386-392. https://doi.org/10.13057/biodiv/d220147

Firdaus, D., Sumardi, I., & Kulsum, Y. (2024). Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(3), 230-243. https://doi.org/10.14421/jiska.2024.9.3.230-243

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1), 32. http://arxiv.org/abs/2312.10997

Geberemeskel, G. A., Debebe, Y. G., & Nguse, N. A. (2019). Antidiabetic effect of fenugreek seed powder solution (Trigonella foenum-graecum L.) on hyperlipidemia in diabetic patients. *Journal of diabetes research*, *2019*(1), 8507453. https://doi.org/10.1155/2019/8507453

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Kartini, K., Jayani, N. I. E., Octaviyanti, N. D., Krisnawan, A. H., & Avanti, C. (2019, December). Standardization of some Indonesian medicinal plants used in "Scientific Jamu". In *IOP Conference Series: Earth and Environmental Science* (Vol. 391, No. 1, p. 012042). IOP Publishing. https://doi.org/10.1088/1755-1315/391/1/012042

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474. http://arxiv.org/abs/2005.11401

Mardiansyah, M. (2016). Ethnobotanical study of herbal medicine in ranggawulung urban forest, subang district, west java, Indonesia. *Biodiversitas Journal of Biological Diversity*. https://doi.org/10.3390/ijerph17103376

Putri, L. S. E., Dasumiati, Kristiyanto, Mardiansyah, Malik, C., Leuvinadrie, L. P., & Mulyono, E. A. (2016). Ethnobotanical study of herbal medicine in Ranggawulung Urban Forest, Subang District, West Java, Indonesia. *Biodiversitas*, *17*(1), 172–176. https://doi.org/10.13057/biodiv/d170125

Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S., & Pasquali, S. (2024, November). Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 608-616). http://arxiv.org/abs/2408.04948

Sholikhah, E. N. (2016). Indonesian medicinal plants as sources of secondary metabolites for pharmaceutical industry. *J Med Sci*, *48*(4), 226-239. https://doi.org/10.19106/jmedsci004804201606

Sianipar, E. A. (2021). The potential of Indonesian traditional herbal medicine as immunomodulatory agents: a review. *International Journal of Pharmaceutical Sciences and Research*, *12*(10), 5229. https://doi.org/10.13040/IJPSR.0975-8232.12(10).5229-37

Sumarni, W., Sudarmin, S., & Sumarti, S. S. (2019, October). The scientification of jamu: A study of Indonesian's traditional medicine. In *Journal of Physics: Conference Series* (Vol. 1321, No. 3, p. 032057). IOP Publishing. https://doi.org/10.1088/1742-6596/1321/3/032057

Woerdenbag, H. J., & Kayser, O. (2014). Jamu: Indonesian traditional herbal medicine towards rational phytopharmacological use. *Journal of herbal medicine*, *4*(2), 51-73. https://doi.org/10.1016/j.hermed.2014.01.002

Zubiaga, A. (2024). Natural language processing in the era of large language models. *Frontiers in artificial intelligence*, *6*, 1350306. https://doi.org/10.3389/frai.2023.1350306.