Optimizing Knowledge Extraction and Knowledge Retrieval in RAG for a Mental Health Chatbot

1st Ega Rizky Setiawan
Department of Electrical and Information Engineering
Gadjah Mada University
Yogyakarta, Indonesia
ega.rizky.setiawan@mail.ugm.ac.id

2nd Bimo Sunarfri Hantono
Department of Electrical and Information Engineering
Gadjah Mada University
Yogyakarta, Indonesia
bhe@ugm.ac.id

3rd Guntur Dharma Putra
Department of Electrical and Information Engineering
Gadjah Mada University
Yogyakarta, Indonesia
gdputra@ugm.ac.id

***Abstract*—The use of chatbots as accessible mental health support tools faces a crucial challenge, the inherent risk of misinformation and "hallucination" in Large Language Models (LLMs). This research aims to address this problem by designing, implementing, and evaluating a Retrieval-Augmented Generation (RAG) architecture built upon a structured Knowledge Graph (KG). The proposed methodology focuses on three main pillars that is designing a RAG architecture that integrates the KG to ensure consistency and accuracy, optimizing the Knowledge Extraction process using an LLM to build the KG from mental health literature, and evaluating various Knowledge Retrieval methods to find the most relevant information. Experimental results show that the RAG-KG architecture significantly outperforms the Naive RAG approach, which is based on document chunks in almost all evaluation metrics. In the evaluation of retrieval methods, the Neighbor Expansion strategy proved to be the most effective overall, surpassing pure vector-based search methods in generating high-quality final answers. This research confirms that the use of a Knowledge Graph, built and accessed via optimized methods, is a robust approach to enhancing chatbot reliability and safety. The result is a system capable of delivering more factual, relevant, and evidence-based mental health information, while simultaneously minimizing the risk of hallucination.**

***Index Terms*—RAG, knowledge graph, knowledge extraction, knowledge retrieval, chatbot, mental health, LLM**

I. Introduction

Mental health is a critical component of individual and societal well-being, yet access to quality care remains a significant challenge globally and in Indonesia [1]. Factors such as social stigma, a shortage of professionals, and high costs create a substantial gap between the need for and the availability of services [2]. Digital solutions, particularly AI-powered chatbots, offer a promising avenue to provide accessible, first-line psychological support [3]. However, standard Large Language Model (LLM) powered chatbots are prone to factual inaccuracies and "hallucinations"—generating plausible but incorrect information. This risk is unacceptable in the sensitive domain of mental health. To mitigate this, the Retrieval-Augmented Generation (RAG) architecture has been proposed, which grounds LLM responses in external knowledge sources [4]. While conventional RAG systems using unstructured text chunks are an improvement, they often struggle to capture the complex, interconnected nature of knowledge in specialized domains like mental health. A Knowledge Graph (KG), which represents information as a network of entities and their relationships, offers a more structured and powerful foundation for reasoning and precise information retrieval [5]. This paper addresses the limitations of existing systems by proposing an optimized RAG architecture based on a Knowledge Graph. Our primary contributions are:

- The design and implementation of a KG-RAG architecture that significantly enhances the factual accuracy and consistency of a mental health chatbot.
- An optimized Knowledge Extraction pipeline using an LLM with few-shot prompting to effectively construct a domain-specific KG from Indonesian mental health literature.
- A comprehensive evaluation of various Knowledge Retrieval strategies, demonstrating that adaptive graph traversal methods outperform standard vector search and baseline RAG.

II. Related Work

The development of conversational agents has evolved through several paradigms. Early systems like ELIZA were rule-based, offering high predictability but suffering from rigidity and poor scalability [6]. With the advent of LLMs, fine-tuning became a popular approach to adapt models to specific domains [7]. While effective at capturing style and specific knowledge, fine-tuning is computationally expensive, requires large, high-quality datasets, and struggles to incorporate new information without complete retraining, while still being susceptible to hallucination.

The RAG framework, introduced by Lewis et al., addresses the issue of static knowledge by combining a retriever with a generator [4]. The retriever fetches relevant text passages from a corpus, which are then used as context by an LLM to generate a factually grounded response. This approach has proven effective in knowledge-intensive NLP tasks.

More recently, researchers have explored replacing unstructured text corpora with Knowledge Graphs as the knowledge source for RAG. A KG models entities (e.g., "Depression") and relationships (e.g., "is treated by") explicitly, enabling more precise retrieval and complex reasoning than is possible

with simple document chunks [5]. The GraphRAG approach, for instance, leverages graph structures and community detection to provide global context for summarization tasks. Our work builds on this foundation by focusing specifically on optimizing the extraction and retrieval mechanisms for a question-answering chatbot in the mental health domain.

III. System Model

Our proposed system consists of two core pipelines: the Knowledge Graph Construction Pipeline, responsible for offline data processing, and the Retrieval and Generation Pipeline, responsible for online query handling.

A. Knowledge Graph Construction Pipeline

The KG is built from a corpus of Indonesian mental health documents, including technical guides from the Ministry of Health and articles from trusted sources. The process, illustrated in Fig. 1, involves several automated steps.

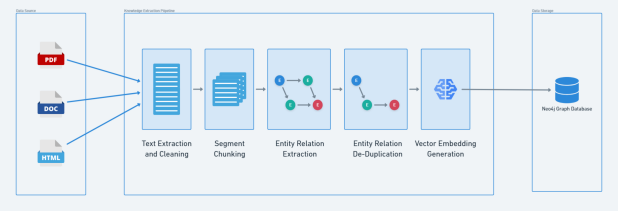

Fig. 1. Knowledge Extraction pipeline.

- Text Extraction & Chunking: Text is extracted from source files (PDF, DOCX, HTML). To manage long documents and preserve context, we employ structural chunking, where documents are split based on semantic sections (e.g., chapters or headings) rather than fixed-size blocks. This approach aims to avoid incomplete context due to defect in document slicing. Incompleteness can lead to LLM failure in understanding a concept in document.
- Entity & Relation Extraction: We use a powerful LLM (Google Gemini 2.5 Flash) to extract entities (e.g., Gangguan Mental, Terapi Psikologis) and their relationships from each text chunk. This is guided by a designed few-shot prompt that provides the LLM with instructions, a predefined ontology of entity and relation types, and examples of input and desired result to ensure high-quality, structured output in JSON format.
- Entity Resolution & Embedding: Extracted entities are deduplicated and merged to maintain consistency across the KG. The descriptions of duplicate entities are combined. Following this, a vector embedding is generated for each unique entity using the Google text-embedding-001 model to enable semantic search capabilities.
- Graph Loading: The final entities (as nodes) and relations (as edges) are loaded into a Neo4j graph database, which supports graph-based queries using Cypher, text-based search, and native vector search.

B. Retrieval and Generation

When a user submits a query, the online pipeline illustrated in Fig. 2 processes it to generate a grounded answer.

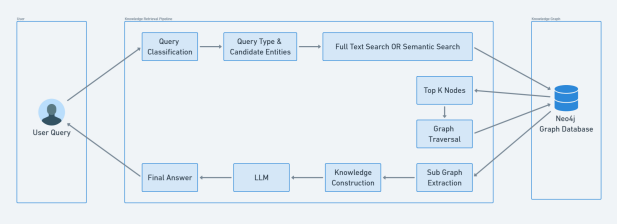

Fig. 2. Knowledge Retrieval and generation pipeline.

- Query Classification: The user's query is first analyzed by an LLM to classify its intent as either an entity_query (seeking information about a single concept, e.g., "What is depression?") or a path_query (exploring the relationship between multiple concepts, e.g., "How does CBT treat anxiety?"). The LLM also extracts key entities from the query used for search keyword.
- Hybrid Search: The system identifies initial candidate nodes in the KG using a hybrid search strategy. It first attempts a fast full-text search for text matches. If unsuccessful, it falls back to a semantic vector search to find the most conceptually similar nodes.
- Adaptive Graph Traversal: Based on the query classification, an appropriate graph traversal algorithm is selected to gather relevant context: For entity_query, a one-hop Neighbor Ex pansion is used to retrieve the central node and its immediate neighbors, effectively gathering its defining attributes and direct relationships. For path_query, an N-Shortest Path algorithm is used to find the most direct connection(s) between the multiple entities identified in the query.
- Context Transformation & Generation: The retrieved subgraph (nodes and edges) is transformed into a structured, human-readable text format. This context, along with the original query, is passed to the LLM, which generates the final, evidence-based answer for the user.

IV. Performance Evaluation

A. Experimental Setup

To evaluate our system, we created a synthetic dataset of 433 question-answer pairs derived from our knowledge corpus (Knowledge Graph and original documents) using OpenAI o4-mini, ensuring the evaluation model was different from the one used in our pipeline to avoid bias. We compared a Naive RAG system (using standard document chunks and vector search) as a baseline against four configurations of our KG-RAG system, which varied the search and traversal methods:

- Default: Hybrid search with automatic traversal selection.
- Vector Search: Pure vector search for initial nodes.
- Neighbor Expansion: Forced neighbor expansion traversal.
- N-Shortest Path: Forced shortest path traversal.

B. Retriever Performance

We evaluated the quality of the context retrieved using Precision, Recall, Mean Reciprocal Rank (MRR), and Hit Ratio.

TABLE I
RETRIEVER EVALUATION

Treatment	Precision	Recall	MRR	Hit Ratio
Default	0.1530	0.8203	0.3758	0.9053
Vector Search	0.0603	0.5929	0.1737	0.6744
Neighbor Expansion	0.1123	0.8868	0.3191	0.9584
N-Shortest Path	0.1245	0.5370	0.2788	0.6097

The results in Table I show a clear trade-off. The Default method achieved the highest precision and MRR, indicating it is best at returning highly relevant results at the top of the list. However, Neighbor Expansion excelled in recall and hit ratio, demonstrating its superior ability to find all relevant information, even if it includes some noise. Pure Vector Search performed poorly across all metrics.

C. End-to-End System Performance

We used the RAGAS framework to evaluate the quality of the final generated answers based on correctness, relevance, and faithfulness to the retrieved context [8].

TABLE II
FINAL ANSWER EVALUATION

Treatment	Similarity	Correctness	Relevancy	Faithfulness
Naive RAG	0.9048	0.3391	0.6996	0.8993
Default	0.9242	0.4155	0.8624	0.9435
Vector Search	0.9115	0.3895	0.7387	0.8902
Neighbor Expansion	0.9248	0.4246	0.8723	0.9428
N-Shortest Path	0.9033	0.3936	0.7900	0.4614

As shown in Table II, the KG-RAG methods significantly outperformed the Naive RAG baseline. The Neighbor Expansion strategy emerged as the overall best performer, demonstrating substantial improvements in key areas. Specifically, it increased the Correctness score by 25.2% (from 0.3391 to 0.4246) and the Relevancy score by 24.7% (from 0.6996 to 0.8723) compared to the baseline. Conversely, the extremely low Faithfulness score for N-Shortest Path (0.4614) highlights a critical finding: if the retrieval process fails to find a relevant path, the LLM is highly likely to hallucinate an answer, even if that answer seems plausible. This underscores the vital importance of a robust and contextually appropriate retrieval mechanism for generating reliable responses.

V. CONCLUSION

This research successfully designed, implemented, and validated an optimized RAG architecture using a Knowledge Graph for a mental health chatbot. Our findings demonstrate that this approach is superior to conventional RAG systems based on unstructured documents. By leveraging a structured KG and an adaptive retrieval strategy, our system produces answers that are more accurate, relevant, and faithful to the source knowledge, thereby minimizing the risk of harmful misinformation. The Neighbor Expansion traversal strategy proved to be the most effective method for retrieving comprehensive context that leads to high-quality final answers. Future work could explore more advanced graph structures and evaluate the clinical impact of the chatbot with human users.

REFERENCES

[1] Center for Reproductive Health, University of Queensland, and Johns Hopkins Bloomberg School of Public Health, "Indonesia - national adolescent mental health survey (i-namhs): Laporan penelitian," Pusat Kesehatan Reproduksi, Yogyakarta, Indonesia, Tech. Rep., 2022, edisi pertama, Oktober 2022.

[2] World Health Organization, *World mental health report: transforming mental health for all*. Geneva: World Health Organization, 2022, licence: CC BY-NC-SA 3.0 IGO. Available: https://www.who.int/publications/i/item/9789240049338.

[3] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Ment Health*, vol. 4, no. 2, p. e19, Jun 2017. [Online]. Available: http://mental.jmir.org/2017/2/e19/

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[5] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," 2025. [Online]. Available: https://arxiv.org/abs/2404.16130

[6] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[7] H. Yu and S. McGuinness, "An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system," *Journal of Medical Artificial Intelligence*, pp. 1–16, 2024.

[8] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.