JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 8, No. 4. September 2026 DOI: https://doi.org/10.34288/jri.v8i4.650
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

IMPLEMENTATION OF DISASTER MITIGATION CHATBOT USING GRAPH RETRIEVAL-AUGMENTED GENERATION (GraphRAG) FOR BPBD KABUPATEN BOGOR

Anastia Firyal Nisrina[-1], Foni Agus Setiawan[-2], Gibtha Fitri Laxmi[-3]

Department of Informatics Engineering, Faculty of Engineering
Universitas Ibn Khaldun Bogor, Bogor, Indonesia
astinisrina02@gmail.com [1)], gibtha@uika-bogor.ac.id [3)]

Research Center for Data and Information Science
National Research and Innovation Agency, Bandung, Indonesia
foni001@brin.go.id [2)]

Abstract

Kabupaten Bogor is among the highest disaster-risk regencies in West Java, having recorded 1,826 disaster events throughout 2025, while official mitigation guidelines remain confined to static PDF documents that are difficult to consult during emergencies. To address this gap, this study implements a disaster mitigation chatbot built on Graph Retrieval-Augmented Generation (GraphRAG), pairing Neo4j Desktop as a local graph database with Gemma3:4b (served through Ollama) as the language model, orchestrated via LangChain and exposed through a Flask web interface. The knowledge graph was built from 11 source documents provided by BPBD Kabupaten Bogor—Disaster Preparedness Pocket Books together with supplementary disaster-information documents—spanning 8 disaster types, yielding 483 nodes and 754 relationships. Each entity node carries three properties (name, description, and disaster phase) to enable more precise retrieval. System performance was assessed with the RAGAS v0.4.3 framework over a 99-question dataset over three trials, yielding average Faithfulness, Answer Relevancy, and Context Precision scores of 0.8764, 0.6773, and 0.8216, respectively. These results indicate promising contextual grounding and retrieval performance, while the lower Answer Relevancy score highlights the need for further improvement in aligning generated answers with user questions. The findings demonstrate the technical feasibility of a locally hosted GraphRAG prototype for retrieving disaster-mitigation information from BPBD Kabupaten Bogor documents. However, the evaluation does not establish operational readiness for emergency use, as comparison with conventional RAG, expert assessment, and validation with intended users remain necessary.

Keywords: Disaster Mitigation; Chatbot; Knowledge Graph; GraphRAG; RAGAS

Abstrak

Kabupaten Bogor merupakan salah satu wilayah dengan risiko bencana tertinggi di Jawa Barat, dengan 1.826 kejadian bencana tercatat sepanjang tahun 2025, sementara panduan mitigasi resmi masih berbentuk dokumen PDF statis yang sulit diakses saat kondisi darurat. Untuk mengatasi kesenjangan ini, penelitian ini mengimplementasikan chatbot mitigasi bencana berbasis Graph Retrieval-Augmented Generation (GraphRAG), yang memadukan Neo4j Desktop sebagai basis data graf lokal dengan Gemma3:4b (dijalankan melalui Ollama) sebagai model bahasa, diorkestrasi menggunakan LangChain dan diakses melalui antarmuka web Flask. Graf pengetahuan dibangun dari 11 dokumen sumber yang disediakan oleh BPBD Kabupaten Bogor—Buku Saku Siaga Bencana beserta dokumen informasi bencana pendukung—yang mencakup 8 jenis bencana, menghasilkan 483 node dan 754 relasi. Setiap node entitas memiliki tiga properti (nama, deskripsi, dan fase bencana) untuk mendukung pengambilan informasi yang lebih presisi. Kinerja sistem dievaluasi menggunakan kerangka kerja RAGAS v0.4.3 terhadap 99 pertanyaan melalui tiga kali pengujian. Hasilnya menunjukkan skor rata-rata untuk Faithfulness, Answer Relevancy, dan Context Precision masing-masing sebesar 0,8764, 0,6773, dan 0,8216. Angka-angka ini mengindikasikan kinerja yang menjanjikan dalam hal landasan kontekstual dan pengambilan informasi (retrieval). Di sisi lain, skor Answer Relevancy yang lebih

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

rendah menunjukkan perlunya peningkatan lebih lanjut agar jawaban yang dihasilkan lebih selaras dengan pertanyaan pengguna. Temuan ini membuktikan kelayakan teknis dari purwarupa GraphRAG yang di-hosting secara lokal untuk mengambil informasi mitigasi bencana dari dokumen-dokumen BPBD Kabupaten Bogor. Akan tetapi, evaluasi ini belum menunjukkan kesiapan operasional untuk penggunaan dalam situasi darurat, karena masih diperlukan perbandingan dengan sistem RAG konvensional, penilaian oleh pakar, serta validasi dengan pengguna yang dituju.

Kata kunci: Mitigasi Bencana; Chatbot; Knowledge Graph; GraphRAG; RAGAS

INTRODUCTION

Indonesia ranks among the world's most disaster-vulnerable nations. Sitting at the junction of three major tectonic plates—the Eurasian, Indo-Australian, and Pacific plates—the country is exposed to a wide range of natural hazards, from earthquakes, volcanic eruptions, and tsunamis to landslides, floods, and strong winds. The National Disaster Management Agency (BNPB) reports more than 3,000 disaster events annually, a figure that continues to climb alongside global climate change (BPBD Kabupaten Bogor, 2025).

As one of West Java's highest-risk regencies, Kabupaten Bogor spans roughly 2,710.62 km^2, stretching from lowlands in the north to mountainous terrain in the south that includes the slopes of Mount Salak, Mount Gede, and Mount Pangrango. BPBD Kabupaten Bogor data show 2,160 disaster events in 2023 alone, resulting in 29 fatalities, 762 evacuees, and 3,910 damaged houses (BPBD Kabupaten Bogor, 2025). More recent figures through early 2026 put the 2025 total at 1,826 events, dominated by strong winds (919), landslides (587), and floods (163) (BPBD Kabupaten Bogor, 2025).

Official mitigation documents, including the BNPB Disaster Pocket Book Edition 2023, are distributed as dense, static PDF files, leaving communities and field officers to manually search thousands of lines of text during emergencies (BNPB, 2023). Despite the availability of official disaster-mitigation guidelines, retrieving actionable information from document-based resources can require users to identify the relevant hazard, locate the appropriate procedural guidance, and distinguish recommendations according to the disaster phase. These requirements create an information-retrieval challenge for conversational systems, particularly when a question depends on relationships among disaster types, mitigation actions, locations, and temporal phases. However, the extent to which these difficulties affect communities and field officers in Kabupaten Bogor has not been empirically established in this study.

Advances in artificial intelligence—particularly Large Language Models (LLMs)—open a path toward chatbots capable of interacting with disaster information conversationally, a trend documented in a recent bibliometric review of AI applications in disaster management (Wibowo et al., 2025). Applied directly, however, LLMs remain prone to factual inaccuracy (hallucination), a weakness that can carry serious consequences in disaster-management contexts (Zhao et al., 2023).

Retrieval-Augmented Generation (RAG) was developed precisely to let LLMs draw on validated external knowledge bases rather than relying solely on parametric memory (Gao et al., 2024). Conventional RAG provides a means of grounding language-model responses in external documents, but retrieval based primarily on semantic similarity may not consistently preserve or retrieve the relationships required by multi-entity, phase-specific questions. Graph-based retrieval offers an alternative by explicitly representing entities and their relationships. Nevertheless, the benefit of this approach for disaster-mitigation information cannot be assumed from its architecture alone and requires comparison under equivalent experimental conditions.

Accordingly, this study implements GraphRAG using official BPBD Kabupaten Bogor documents on a Neo4j Desktop graph database, representing knowledge as explicit entities and relationships to enable semantic search alongside more accurate relational traversal (Edge et al., 2025). System performance is then validated using the RAGAS framework against the disaster mitigation domain of Kabupaten Bogor (Es et al., 2024). The study investigates the feasibility of this approach and its retrieval and generation performance, rather than claiming validated superiority over conventional RAG or readiness for operational emergency deployment.

This study has three main contributions: (1) Disaster-domain graph representation: A knowledge graph constructed from 11 official

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.650
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

BPBD documents, with :Entitas and :Dokumen node labels, entity descriptions, disaster-phase properties, and document-to-entity relationships; (2) Phase-aware retrieval pipeline: An implemented retrieval mechanism that classifies questions, maps disaster-related keywords to source documents, retrieves connected entities and relationships using Cypher, and filters the results according to the detected disaster phase; and (3) Empirical prototype evaluation: An evaluation of the implemented system using 99 questions, three trials, and three metrics, with analysis of grounding, retrieval precision, and answer relevance.

RELATED WORK

A number of studies have explored chatbot-based approaches to disaster risk communication. Wickramasinghe et al. (2024) built a chatbot to disseminate landslide-resilient construction guidelines for high-risk areas in Sri Lanka, fine-tuning PHI-3 and LLAMA-3 on a 5,000-entry Question-and-Answer dataset via the Unsloth framework with Low-Rank Adaptation (LoRA); their comparison found PHI-3 generalized better, producing more contextually relevant and detailed responses, though the system remains tied to a static fine-tuned dataset rather than a dynamically queryable knowledge structure.

Urbanelli et al. (2024) introduced the ERMES Chatbot, a Telegram-based tool designed through user-centred design to support two-way communication among citizens, field responders, and control rooms across the prevention, response, and recovery phases of emergency management; field trials across four European countries yielded a System Usability Scale score of 64.3, with the tool proving most useful for crowdsourced reporting and broadcast communication, though it does not incorporate retrieval-augmented generation to ground its responses in a structured knowledge base.

Within the Indonesian context, Sintawati et al. (2025) reviewed the role of AI in flood disaster mitigation, examining its use across pre-disaster early warning, emergency response, and post-disaster rehabilitation. Their review found that AI-powered conversational systems—including chatbots for automated evacuation guidance—are emerging as a key tool for strengthening public readiness and inter-agency coordination, and called for more structured, knowledge-grounded implementations within Indonesian disaster management. The present study responds directly to that call, applying a structured Knowledge Graph as the retrieval backbone for a disaster mitigation chatbot validated against official BPBD Kabupaten Bogor documents.

More recent Indonesian implementations illustrate the range of platforms explored for this purpose: Wati and Nugrahantoro (2026) built a WhatsApp-based disaster mitigation chatbot using n8n workflow automation paired with an AI Agent, while Hikmawati (2024) developed the ASIAB chatbot to teach disaster-preparedness concepts to primary-school students, validated through the ADDIE model. Neither system, however, grounds its responses in a structured, queryable knowledge base—the gap the present GraphRAG-based system addresses.

Relative to these systems, the present research specifically targets the risk of hallucination in technical mitigation guidance by grounding chatbot responses in a structured knowledge graph rather than a fine-tuned dataset or a rule-based dialogue flow. Gao et al. (2024) formalized the GraphRAG taxonomy into G-Indexing, G-Retrieval, and G-Generation stages, demonstrating graph-based approaches' advantage in cross-document entity resolution. Edge et al. (2025) proposed the Document Knowledge Graph (DKG) to preserve hierarchical document structure, improving Context Relevance. Es et al. (2024) introduced RAGAS at EACL 2024, enabling automated RAG evaluation without manually annotated references. Neo4j's LangChain integration (Neo4j, Inc., 2024) supports hybrid search combining vector storage with Cypher query execution. Zhang et al. (2024) further showed that domain-specific graph schemas yield markedly better accuracy than generic Knowledge Graphs, while Pan et al. (2025) emphasized that unifying LLMs with structured Knowledge Graphs to curb hallucination remains an active research direction. This direction is reinforced by work explicitly combining knowledge graphs with LLMs for disaster and emergency contexts: Chen et al. (2024) proposed a knowledge-graph-augmented LLM framework for emergency decision-making, and Xie et al. (2025) outlined a knowledge-graph-driven AI framework for disaster prevention, mitigation, and emergency response, both underscoring the value of structured graph representations for grounding LLM outputs in high-stakes domains.

METHODOLOGY

This research applies the methodology presented in Figure 1, from data collection activities to performance evaluation using RAGAS.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

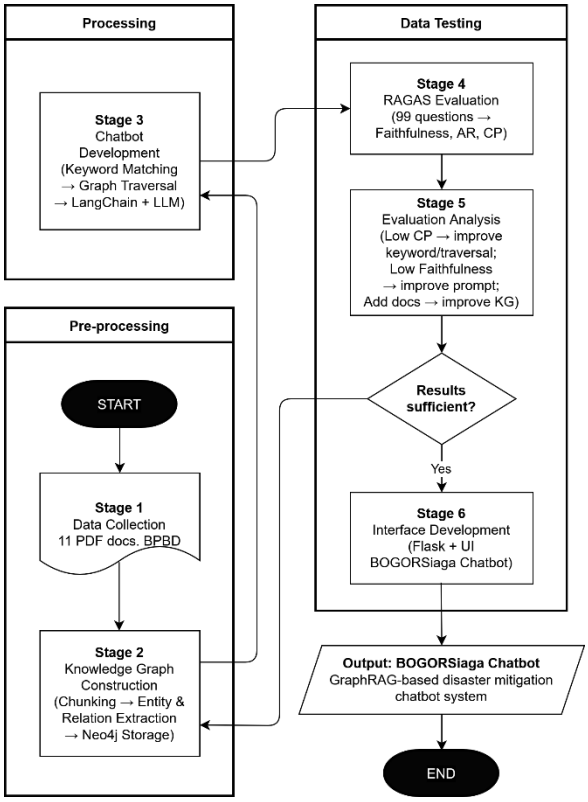


Figure 1. Research methodology flowchart

System Architecture and Tools

This research relies on a combination of hardware and software tailored to local computing needs. The system runs on Windows 10 (64-bit) with an Intel Core i3-1215U processor (8 cores, 1.20 GHz), 12 GB DDR4 RAM, and a 512 GB NVMe SSD. Core software components include Neo4j Desktop 5.x (Community Edition) as the graph database—chosen for its native property-graph model, which prior Indonesian work has shown to represent connected, relationally rich data more intuitively than an equivalent relational schema (Sholeh et al., 2020)—LangChain v0.2.x as the orchestration framework, Gemma3:4b via Ollama as the local language model, text-embedding-3-small (OpenAI API) for embeddings, RAGAS v0.4.3 for evaluation, Flask 3.x for the web interface, and PyMuPDF v1.24 for PDF parsing. Like other modern LLMs, Gemma3:4b is built on the transformer architecture (Vaswani et al., 2017), which underpins its language understanding and generation capabilities.

Data Collection

The primary dataset comprises 8 official BPBD Kabupaten Bogor documents (Buku Saku Siaga Bencana; BPBD Kabupaten Bogor, 2023) covering 8 disaster types—Strong Winds, Floods, Landslides, Earthquakes, Volcanic Eruptions, Drought, Forest and Land Fires, and Tsunamis—supplemented by 3 additional BPBD documents covering general disaster preparedness, emergency contact directories, and cross-disaster procedural guidance, for 11 source documents in total. All documents were obtained from the verified official BPBD Kabupaten Bogor portal.

Knowledge Graph Construction (G-Indexing)

PDF documents were extracted with PyMuPDF while preserving page structure. Preprocessing steps included raw text extraction with special handling for tables and bullet lists, removal of PDF artifacts such as repeated headers and footers, segmentation following the chapter-section-subsection hierarchy, and semantic chunking at 512 tokens with 10–15% overlap to preserve continuity of meaning.

G-Indexing relied on a two-label node schema: :Entitas for every extracted entity—objects, locations, actions, institutions, and domain concepts—and :Dokumen for the 11 source PDF documents. Each :Entitas node carries three properties: name (normalized to lowercase to avoid duplication from capitalization variants), description (a brief explanation grounded in the source sentence, left blank when the source text does not explicitly explain the entity, to guard against hallucination), and phase (a Pre-/During-/Post-disaster label for phase-specific actions, left blank for phase-agnostic entities). Adding explicit description and phase properties at this stage proved critical to answer quality, since graph-based retrieval could then supply factual grounding and phase context directly, rather than leaving the generation LLM to reason about it implicitly. Unlike fixed-schema approaches, relationship types were generated dynamically by Gemma3:4b during extraction, yielding 79 unique relation types tailored to each document's context; the most dominant include MEMILIKI_ENTITAS/ MEMBAHAS (document-to-entity structural links), MEMBUTUHKAN, TERJADI_DI, and BERDAMPAK_PADA. Using few-shot prompting, Gemma3:4b identified entities and declared relationships as triplets (Subject–Relation–Object), each automatically validated before being stored in Neo4j Desktop. Ingesting all 11 source documents took approximately 221.6 minutes (~3.7 hours) on local hardware (12 GB RAM, CPU-only) via Ollama, reflecting the computational cost of LLM-based entity and relation extraction, including generation

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.650
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

of the description and phase properties for every entity.

GraphRAG Pipeline (G-Retrieval & G-Generation)

The retrieval pipeline was built in Python and Flask, with LangChain's Neo4jGraph and OllamaLLM wrappers providing access to the graph database and language model, and is organized around three main components. The Question Classification Module first classifies user questions via keyword-phrase matching into five categories—definition/theory, location/area, impact/hazard, prohibition, or procedural—and detects the temporal disaster phase (pre-, during-, or post-disaster) from keywords in the question. The Graph Retrieval Module then maps disaster-specific keywords to their corresponding source documents through a keyword-to-document dictionary, executes Cypher queries to retrieve all entities connected to the matched document(s) along with their inter-entity relations, and prioritizes and filters those entities by matching their phase property against the question's detected phase, which reduces cross-phase contamination in the assembled context. Finally, the Answer Generation Module composes the phase-labeled context together with a category-specific instruction into a prompt that guides Gemma3:4b toward a natural-language answer grounded in that context.

Evaluation Dataset

The 99 questions were formulated from the content of the 11 source documents so that every disaster type is covered across the pre-, during-, and post-disaster phases. The set covers the 8 disaster types plus a general BPBD category of institutional questions (for example the role of BPBD, assembly points, evacuation routes, and emergency contacts). By categories assigned by the question classification module, 76 are procedural, 15 definition/theory, 6 impact/hazard, 1 location/area, and 1 prohibition question.

The reference answer for each question was written independently from the source documents, without consulting the retrieved context or the chatbot's answers. In addition to RAGAS, we manually compared every chatbot answer with the source documents and rated it as Fully Consistent, Mostly Consistent, or Inconsistent. The questions were not yet reviewed by external domain experts; this is acknowledged as a limitation. Fourteen of the 99 questions address topics that are not explicitly covered in the 11 source documents; for these, the expected behavior is a fallback notification that the information is unavailable.

Evaluation Using RAGAS Framework

Testing relied on RAGAS v0.4.3 over a 99-question evaluation dataset spanning 8 disaster types, each represented by questions covering the pre-disaster, during-disaster, and post-disaster phases, with ground truth drawn from BPBD Kabupaten Bogor's Disaster Preparedness Pocket Books. Three evaluation metrics were applied: Faithfulness, Answer Relevancy, and Context Precision (Patronus AI, 2024). The 0.80 value used throughout this paper is a target threshold defined by the authors as a reference point for comparing the three metrics; it is not a validated standard for operational deployment.

Faithfulness measures the proportion of claims in an answer that can be verified against the retrieved context, with scores approaching 1.0 indicating an answer fully grounded in that context.

$$F = \frac{|V|}{|S|} \quad (1)$$

where |V| is the number of verified claims and |S| is the total number of claims in the answer.

Answer Relevancy captures how well a generated answer addresses the intent behind the user's question, computed as the average cosine similarity between the original question and N questions generated from the answer. Because RAGAS v0.4.3 produces a NaN error with local embeddings, this metric was instead calculated manually using text-embedding-3-small.

$$AR = \frac{1}{N}\sum_{i=1}^{N} \cos(q, Q_i) \quad (2)$$

where q is the original question and Qi are the questions generated from the answer.

Context Precision checks whether the system retrieves only truly relevant graph nodes, by comparing the retrieved context against an independently-authored reference. To keep this comparison methodologically sound, the reference used here was written directly from the 11 source documents, without consulting the retrieved context or the chatbot's own answers.

$$CP = \frac{|E|}{|C|} \quad (3)$$

where |E| is the number of relevant contexts retrieved and |C| is the total number of contexts retrieved.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 | E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.650
JURNAL RISET INFORMATIKA
Vol. 8, No. 4. September 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

RESULTS AND DISCUSSION

Knowledge Graph Construction Results

G-Indexing over the 11 BPBD Kabupaten Bogor source documents successfully produced a disaster mitigation knowledge graph stored in Neo4j Desktop; Table 1 summarizes the resulting construction statistics.

Table 1. Knowledge graph construction statistics

Parameter	Value
Total Nodes	483
Total Relations	754
Source Documents	11 PDF documents
Disaster Types Covered	8 types
Ingest Time (11 docs)	~221.6 minutes (~3.7 hours)
Average Nodes/Document	~44 nodes
Average Relations/Document	~68 relations

The resulting knowledge graph uses two node labels: :Entitas (472 nodes), covering every extracted entity—objects, locations, actions, institutions, and concepts within the disaster mitigation domain—and :Dokumen (11 nodes), representing the 11 source PDF documents ingested into the graph. Each :Entitas node carries name, description, and disaster-phase properties (pre-, during-, or post-disaster, left blank when phase-agnostic); of the 472 :Entitas nodes, 349 (73.9%) have a populated phase property (156 pre-disaster, 161 during-disaster, 32 post-disaster). Node-label distribution is presented in Table 2.

Table 2 breaks down the top relation types in the Knowledge Graph, totaling 754 relations across 79 unique types. MEMBAHAS/MEMILIKI_ENTITAS relations dominate, representing the structural link between each :Dokumen node and the :Entitas nodes it discusses, while semantic relations such as MEMBUTUHKAN, BERDAMPAK_PADA, and TERJADI_DI capture the operational, spatial, and causal connections between entities within the disaster mitigation domain. Figure 2 visualizes the graph structure in Neo4j Desktop, showing the dense connectivity between :Dokumen and :Entitas nodes through MEMBAHAS/MEMILIKI_ENTITAS relations alongside the denser semantic sub-networks among mitigation-domain entities.

Table 2. Relation type distribution

Relation Type	Count	Meaning
MEMBAHAS	504	Document discusses entity
MEMBUTUHKAN	60	Entity requires another entity
TERJADI_DI	52	Entity or event occurs at location
BERDAMPAK_PADA	38	Entity impacts another entity
MELIBATKAN	49	Entity involves other parties/objects
Others (83 types)	226	Other context-specific relation types
Total	**754**	

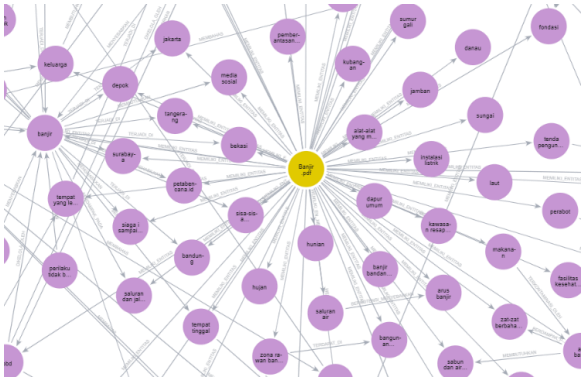


Figure 2. Screenshot of the knowledge graph visualization for BPBD Kabupaten Bogor's disaster mitigation domain using Neo4j Desktop.

GraphRAG Pipeline Implementation

The GraphRAG pipeline was successfully implemented locally through three integrated components. The Question Classification Module correctly sorted questions into five categories—definition/theory, location/area, impact/hazard, prohibition, and procedural—and detected the temporal disaster phase from question keywords. The Graph Retrieval Module then executed Cypher queries to retrieve entities linked to keyword-matched source documents, filtering and prioritizing results by phase relevance; retrieval relies entirely on graph traversal, with no separate vector similarity search component involved.

Total answer-generation time across the 99 evaluation questions initially came to roughly 460 minutes, averaging ~279 seconds per question—a consequence of a retrieval strategy that re-read source documents directly for every query, running on Gemma3:4b locally with 12 GB RAM and no GPU acceleration, compounded by the relatively large context sizes that document-based retrieval produces (up to 25 entities per query, each carrying a description and phase label). As a post-defense refinement, the retrieval stage was switched from re-reading source documents to indexed queries against the Neo4j graph database, cutting the

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.650
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

average response time to roughly 15–20 seconds per question with no meaningful drop in RAGAS scores. The system also issues fallback notifications when requested data cannot be found in the knowledge graph. A further, SQLite-based caching layer was subsequently added: questions identical or highly similar to previously processed ones are now answered directly from cache within seconds, while genuinely new questions still require the full ~15–20-second GraphRAG pipeline described above.

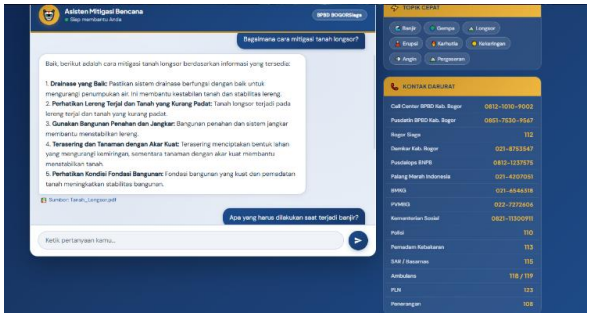


Figure 3. Disaster mitigation chatbot web interface for BPBD Kabupaten Bogor

RAGAS Evaluation Results

Evaluation was carried out over three independent trials using the same 99-question dataset to check for consistency and reliability. The evaluation dataset is shared through the National Scientific Repository (Repositori Ilmiah Nasional) operated by BRIN (Setiawan et al., 2026). Table 3 reports the complete results across all three trials, computed against the final, independently-authored reference described in the Methodology section.

Table 3. RAGAS evaluation results across three times trials

Trial	Faithfulness	Ans. Relevancy	Ctx. Precision
1	0.8752	0.6785	0.8283
2	0.8564	0.6778	0.8283
3	0.8976	0.6757	0.8081
Average	0.8764	0.6773	0.8216

Discussion

The evaluation results indicate different levels of performance across the three metrics. The average Faithfulness score of 0.8764 suggests that generated answers were generally consistent with the retrieved context under the applied evaluation procedure. The Context Precision score of 0.8216 indicates promising retrieval performance against the independently prepared references. However, the Answer Relevancy score of 0.6773 identifies a remaining limitation in the alignment between generated answers and the original questions. These results should therefore be interpreted as evidence of the technical feasibility of the proposed prototype rather than proof of comprehensive reliability or operational readiness. However, in recent studies, these evaluation scores are compared against a threshold value that indicates the system's readiness for implementation in the production phase (acceptable deployment/ production readiness); see (Al-Kadi, 2026), (Chen, 2026), and (Suryanarayana, 2026).

An average Faithfulness of 0.8764 indicates that roughly 87% of the chatbot's answer claims can be verified against the retrieved context, clearing the 0.80 threshold despite Gemma3:4b running locally on just 12 GB RAM with no GPU acceleration. This result reflects a combination of retrieval-level refinements (explicit entity description and phase properties, document-based retrieval, phase-aware filtering) and prompt-level instructions (explicit prohibitions against copying raw context formatting or mixing information across disaster phases). Answer Relevancy, at 0.6773, falls short of the 0.80 threshold but remains reasonable for a 4B-parameter local model. It was computed manually using Formula (2) with N=3 questions generated from each answer using gpt-4.1-nano and then compared via cosine similarity (text-embedding-3-small) against the original question, working around a NaN bug (an AttributeError on the embed_query method) that RAGAS v0.4.3 produces with the local embeddings wrapper.

Context Precision of 0.8216 stands out as one of the study's more significant results, confirming that the GraphRAG pipeline retrieves targeted context more than 82% of the time and clears the 0.80 threshold. An important methodological lesson surfaced during evaluation: early iterations formatted the reference identically to the retrieved context to keep comparison granularity aligned, an approach later found to be methodologically unsound, since Context Precision is meant to compare two independent sources—making them identical yields a score that looks high but carries little information. The final reference was subsequently rewritten independently, drawn directly from the 11 source documents without consulting the retrieved context or chatbot answers, producing the more methodologically defensible score of 0.8216 reported above.

Faithfulness (0.8764) and Context Precision (0.8216) both clearing the 0.80 threshold indicates that retrieval and generation are now performing

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

reliably in tandem, a marked contrast to earlier iterations of this system where Faithfulness trailed well behind Context Precision. This gain stems from a combination of retrieval- and prompt-level refinements: explicit description and phase properties added to each entity during G-Indexing supply factual grounding and temporal context directly. Document-based retrieval surfaces the full set of relevant entities rather than only those whose names happen to match question keywords and prompt instructions explicitly forbid copying raw context formatting or blending information across disaster phases. The remaining gap sits in Answer Relevancy (0.6773), still below the 0.80 threshold because this manual proxy metric compares single generated questions against the original one. The chatbot's deliberately comprehensive, multi-point answers can lower cosine similarity relative to narrower, single-focus answers, even when those longer answers remain accurate and well-grounded. Unlike earlier iterations of the system, then, generation-stage groundedness is no longer the primary bottleneck. Instead, Gemma3:4b's limited capacity (4 billion parameters, running on a 12 GB RAM machine with no GPU acceleration) continues to shape the precision of relevance matching under the manual Answer Relevancy proxy, even after a subsequent retrieval-level optimization brought end-to-end response latency down from ~279 seconds to roughly 15–20 seconds per question.

Set against related disaster-domain conversational systems, these results reveal a different set of trade-offs. Wickramasinghe et al. (2024) fine-tuned PHI-3 and LLAMA-3 on a static 5,000-entry Question-and-Answer dataset for landslide-resilient construction guidance; effective within that fixed scope, the approach cannot absorb new disaster information without costly re-fine-tuning, whereas the knowledge graph used here can simply be extended through incremental ingestion. The graph-based architecture supports the potential incorporation of additional disaster-mitigation documents without requiring model fine-tuning, provided that the new documents can be processed, validated, and integrated into the existing graph. However, the present study did not experimentally evaluate incremental-update time, update accuracy, or the cost of incorporating new documents. The efficiency and reliability of this updating process therefore remain subjects for future investigation.

Urbanelli et al.'s ERMES Chatbot (2024) achieved strong field usability (SUS 64.3) for crowdsourced reporting and broadcast communication, but without retrieval-augmented grounding it was never designed to answer detailed procedural or location-specific mitigation questions—precisely the gap this research targets. Within the Indonesian disaster-management context, Sintawati et al. (2025) flagged structured, knowledge-grounded chatbot implementations as a priority need; the present system answers that call directly by pairing a dynamically generated knowledge graph with a locally hosted LLM, all inside Indonesia's official BPBD operational framework. Against the broader GraphRAG literature, the Context Precision of 0.8216 obtained here aligns with the gains Edge et al. (2025) reported for graph-grounded retrieval over conventional chunk-based RAG, reinforcing that explicit entity-relationship structuring is a key driver of retrieval accuracy in knowledge-intensive, safety-critical domains such as disaster mitigation. The evaluation of GraphRAG usage in this study does not include a controlled conventional RAG baseline or validation by intended emergency users; therefore, the results do not establish that GraphRAG is superior to alternative retrieval approaches or suitable for direct use during emergencies.

This study surfaces several key findings. First, a flexible two-label schema (:Entitas and :Dokumen) with dynamically generated relation types—rather than a fixed ontology—let the knowledge graph capture the semantic richness of the 11 source documents, yielding 483 nodes and 754 relations without manual schema engineering; enriching each entity with explicit description and phase properties further let the retrieval stage supply factual, temporally-labeled context directly, closing much of the gap that previously separated retrieval quality from generation faithfulness. Second, Faithfulness (0.8764) and Context Precision (0.8216) both clearing the 0.80 threshold confirms that grounding and factual correctness can be achieved together even with a small, locally-hosted 4B-parameter model, provided the retrieval stage supplies sufficiently structured and explicit context; the remaining gap in Answer Relevancy (0.6773) underscores the continued need to evaluate RAG systems with multiple complementary metrics rather than retrieval accuracy alone. Third, for BPBD Kabupaten Bogor the practical implication is that a fully local, offline-capable chatbot can already deliver well-grounded disaster mitigation guidance without depending on cloud LLM services—a meaningful advantage in emergency conditions where internet connectivity may be unreliable.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

These results should nonetheless be read alongside their hardware constraints. Running entirely on a CPU-only, 12 GB RAM machine necessarily limits both model choice and inference speed. The original ~279-second average response time was addressed with a post-defense optimization that switched the retrieval stage from re-reading source documents to indexed Neo4j queries, cutting the average to roughly 15–20 seconds per question with no meaningful drop in RAGAS scores; a subsequent SQLite-based caching layer now serves previously answered (or closely matching) questions within seconds, though genuinely novel questions still require the full ~15–20-second GraphRAG pipeline. This remaining latency, together with the manual Answer Relevancy proxy's limitations, directly motivates the future work outlined below, particularly evaluating larger, GPU-accelerated models to determine whether Faithfulness and Answer Relevancy can improve further without sacrificing the Context Precision gains already achieved through the graph-based retrieval design.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study implemented a GraphRAG-based disaster-mitigation chatbot using 11 official BPBD Kabupaten Bogor documents covering eight disaster types. The resulting knowledge graph contains 483 nodes and 754 relationships and supports retrieval through entity relationships, source-document context, question classification, and disaster-phase-aware filtering.

Evaluation using 99 questions over three trials produced average Faithfulness, Answer Relevancy, and Context Precision scores of 0.8764, 0.6773, and 0.8216, respectively. These results indicate promising contextual grounding and retrieval performance, while the lower Answer Relevancy score highlights the need to improve the alignment of generated answers with user questions. The findings demonstrate the technical feasibility of the proposed local prototype but do not establish its superiority over conventional RAG or its readiness for operational emergency use.

The principal limitations are the absence of a controlled conventional RAG comparison, the need for further validation of the evaluation dataset and metrics, and the lack of expert and intended-user assessment. These limitations define the next steps required to assess the system's reliability and practical applicability in disaster-mitigation contexts.

Suggestion

Future work should compare the proposed GraphRAG pipeline with conventional RAG under controlled experimental conditions using identical source documents, evaluation questions, language models, and quality metrics. The Answer Relevancy score should be improved through systematic testing of question classification, phase-aware retrieval, context filtering, and answer-generation prompts. The evaluation dataset and reference answers should also be reviewed by qualified disaster-management experts, preferably in collaboration with BPBD Kabupaten Bogor.

Further studies should assess failure cases, potentially unsafe answers, and the system's performance on questions requiring information that is absent from the knowledge graph. Incremental document ingestion, retrieval latency, and the reliability of knowledge-graph updates should be measured experimentally. The knowledge graph could also be cross-validated against national disaster databases such as DIBI (BNPB, 2024) to further improve coverage and reliability. Finally, usability testing with intended users and assessment of offline deployment reliability are necessary before considering operational use in emergency situations.

REFERENCES

Al-Kadi, O., & Al-Zoubi, A. (2026). From issues to insights: RAG-based explanation generation for software issue tracking systems. *Proceedings of the 2026 International Conference on Software Engineering and Artificial Intelligence*, 142–151.

BNPB. (2023). *Buku saku tanggap tangkas tangguh menghadapi bencana.* Jakarta: Badan Nasional Penanggulangan Bencana.

BNPB. (2024). *Data Informasi Bencana Indonesia (DIBI).* Retrieved from https://dibi.bnpb.go.id/

BPBD Kabupaten Bogor. (2023). *Buku saku siaga bencana.* Bogor: Badan Penanggulangan Bencana Daerah Kabupaten Bogor.

BPBD Kabupaten Bogor. (2025). *Rekapitulasi data bencana tahun 2023–2025.* Bogor: BPBD Kabupaten Bogor.

Chen, L., Zhang, X., & Roberts, J. (2026). A comparative evaluation of RAG architectures for cross-domain LLM applications: Design,

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

implementation, and assessment. *Journal of Enterprise Information Systems*, 20(1), 89–112.

Chen, M., Tao, Z., Tang, W., Qin, T., Yang, R., & Zhu, C. (2024). Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction,* 113, 104804.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2025). *From local to global: A Graph RAG approach to query-focused summarization.* arXiv preprint arXiv:2404.16130.

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). *Retrieval-augmented generation for large language models: A survey.* arXiv preprint arXiv:2312.10997.

Hikmawati, N. (2024). ASIAB chatbot for disaster mitigation learning at Madrasah Ibtidaiyah. *Geosfera Indonesia, 9*(2).

Neo4j, Inc. (2024). Integrating Neo4j with LangChain for GraphRAG. *Neo4j Developer Blog.* Retrieved from https://neo4j.com/blog/developer/

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2025). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering.*

Patronus AI. (2024). RAG evaluation metrics: Best practices for evaluating RAG systems. Retrieved from https://www.patronus.ai/llm-testing/rag-evaluation-metrics

Setiawan, F. A., Nisrina, A. F., & Laxmi, G. F. (2026). *Knowledge Graph Dataset for Disaster Mitigation Chatbot of BPBD Kabupaten Bogor.* https://hdl.handle.net/20.500.12690/RIN/YYCQ5P, RIN Dataverse, V1, UNF:6:FBw+mcI+9mqHC4DASfS1FA== [fileUNF]

Sholeh, M., Rachmawati, R. Y., & Susanti, E. (2020). Pemodelan basis data graph dengan Neo4j (studi kasus: Basis data sistem informasi penjualan pada UMKM). *Jurnal Teknologi Informasi dan Terapan, 7*(1), 25–30.

Sintawati, L., Rusmeilantikasari, W., Maulana, R. H., Hotimah, S., & Setiawan, E. (2025). Pemanfaatan Artificial Intelligence (AI) untuk meningkatkan mitigasi bencana banjir di Indonesia. *Triwikrama: Jurnal Multidisiplin Ilmu Sosial, 9*(11).

Suryanarayana, M., Rao, P. V., & Kumar, S. (2026). Real-time intelligent virtual assistant based on retrieval augmented generation (RIVA): Design and metric evaluation. *International Journal of Artificial Intelligence (IJAI)*, 15(2), 245–258.

Urbanelli, A., Frisiello, A., Bruno, L., & Rossi, C. (2024). The ERMES chatbot: A conversational communication tool for improved emergency management and disaster risk reduction. *International Journal of Disaster Risk Reduction, 112*, 104792.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wati, V., & Nugrahantoro, A. (2026). Development of a WhatsApp-based disaster mitigation chatbot using n8n workflow automation. *Elkom: Jurnal Elektronika dan Komputer, 19*(1), 472–482.

Wibowo, A., Amri, I., Surahmat, A., & Rusdah, R. (2025). Leveraging artificial intelligence in disaster management: A comprehensive bibliometric review. *Jàmbá: Journal of Disaster Risk Studies, 17*(1), a1776.

Wickramasinghe, P., Akalanka, H. M. N., Jayaprakash, S., Chathuranga, U. D. S., & Munasinghe, D. S. (2024). Chatbot-assisted disaster preparedness: A tool for educating the public on landslide-resilient building practices. In *Proceedings of the 14th Annual Research Symposium, National Building Research Organisation, Sri Lanka* (pp. 280–292).

Xie, C., Gao, H., Huang, Y., Xue, Z., Xu, C., & Dai, K. (2025). Leveraging the DeepSeek large model: A framework for AI-assisted disaster prevention, mitigation, and emergency response systems. *Earthquake Research Advances, 5*(4), 100378.

Zhang, H., et al. (2024). Knowledge graph-based RAG for disaster response: A domain-specific approach. *International Journal of Disaster Risk Reduction.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). *A survey of large language models.* arXiv preprint arXiv:2303.18223.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License