Comparative Evaluation of RAG and GraphRAG for Open-Ended Question Answering

Jadesola Osinowo
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: j.osinowo@ulster.ac.uk

Abiodun Adebayo
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: a.adebayo@ulster.ac.uk

Sonya Coleman
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: sa.coleman@ulster.ac.uk

Dermot Kerr
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: d.kerr@ulster.ac.uk

Justin Quinn
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: jp.quinn@ulster.ac.uk

***Abstract*—Retrieval-Augmented Generation (RAG) has become the basis in modern question answering (QA) systems, combining Large Language Models (LLMs) with external document retrieval. However, traditional RAG architectures often struggle with retrieving semantically structured or context-rich content, which can impact accuracy and relevance. This paper presents a comparative evaluation of a standard RAG model and a GraphRAG model. The GraphRAG approach combines graph-based document representation within the retrieval pipeline to assess their efficiency on a structured question answering task. This research leverages two modern Ollama-hosted models, Phi-4 and LLaMA 3.2 3B, as the base language models for both retrieval pipelines. Using a custom dataset derived from German coalition policy documents, this research evaluates performance through both lexical and semantic metrics. The results demonstrate that GraphRAG consistently outperforms traditional RAG in semantic alignment and contextual accuracy, particularly when paired with Phi-4. These findings aim to contribute to the growing body of work on hybrid retrieval strategies and support the case for graph-enhanced architectures in long-form QA systems, which are central to advancing structured and knowledge-aware retrieval methods in complex information domains.**

Keywords- RAG; GraphRAG; LLM; GenAI.

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a foundational method to augment large language models (LLMs) with factual grounding, enabling them to access and condition on external knowledge sources at inference time. By retrieving relevant documents and guiding the generation process with this evidence, RAG bridges the strengths of information retrieval systems and generative language models, reducing hallucinations and improving factual accuracy [1]. However, traditional RAG systems typically treat retrieved text as flat sequences, overlooking deeper semantic, structural, or logical relationships within the context. This limitation has been increasingly recognised as a barrier in handling complex or multi-hop queries, especially across dense knowledge domains like legal, policy, or scientific corporations [2].

To address this, recent innovations such as GraphRAG propose a more structured form of context representation. Rather than presenting documents as flat chunks, GraphRAG represents retrieved information as a graph of entities, concepts, and relationships, enhancing reasoning over inter-related content [3]. This graph-based approach has demonstrated empirical benefits in use cases such as Biomedical Question Answering (BQA), complex multi-hop QA, and long context reasoning tasks [4][13] outperforming traditional RAG in both factual alignment and coherence [4][5]. Despite these promising developments, comparative evaluations remain scarce particularly across different LLM backbones and real-world structured datasets.

For example, in a policy-related query requiring evidence from both cybersecurity and economic governance sections, a traditional RAG system retrieves the top-k most similar text chunks independently. In contrast, GraphRAG traverses relational links between entities across documents, enabling multi-hop integration of evidence before generation. This structured traversal allows the model to construct a more coherent and contextually connected response.

This paper evaluates the performance of traditional RAG and GraphRAG architectures across two modern open-source LLMs hosted via Ollama: Phi-4, a compact model optimised for high-context reasoning, and LLaMA 3.2 3B, known for its larger parameter count and general knowledge coverage. Using a custom benchmark of policy-related QA tasks derived from coalition agreement documents, we assess how well each model-architecture pair performs in grounded, retrieval-heavy generation scenarios. To achieve a holistic evaluation, we employ both lexical overlap metrics (BLEU, ROUGE) and semantic alignment scores using RAGAS including metrics like cosine similarity, faithfulness, and context relevance. The results show that GraphRAG consistently outperforms traditional RAG pipelines, particularly when paired with Phi-4, generating answers that are both more faithful to the source materials and better aligned with user queries. The remainder of this paper is structured as follows: Section II describes the methodology and system architecture, Section III outlines the experimental setup and evaluation framework, Section IV presents the performance results and analysis, and Section V concludes with key findings and future research directions.

II. RELATED WORK

Retrieval-Augmented Generation (RAG) has been extensively studied to integrate information retrieval with large language models, enabling systems to leverage external knowledge at inference time. Previous works have demonstrated the potential for RAG in question answering (QA), summarisation, and dialogue tasks, improving factual grounding compared with standalone LLMs [1]. However, limitations such as shallow retrieval granularity and lack of semantic structuring often reduce performance when queries require multi-hop reasoning or relational context [7].

To overcome these limitations, graph-based retrieval methods have emerged. GraphRAG extends a traditional RAG by organising documents into graph structures, where nodes represent entities or semantic units and edges encode relationships such as co-occurrence, hierarchy, or topical linkage [8]. This approach allows retrieval to follow relational paths, yielding context that is both more compact and semantically meaningful. Recent comparative studies highlight GraphRAG's advantage in tasks requiring multi-document reasoning, such as query-based summarisation or knowledge graph QA, while traditional RAG remains competitive on simpler single-hop factoid tasks [9][10].

A. Traditional RAG Structure

Traditional RAG systems follow a relatively linear pipeline, as illustrated in Figure 1:

1) *Document Preprocessing and Chunking:* Source documents are segmented into overlapping text chunks, often based on fixed token or character lengths [1].
2) *Indexing and Retrieval:* A retriever (sparse, dense, or hybrid) determines the top-*k* chunks most relevant to a query. Dense retrieval methods such as sentence embeddings have become standard, improving semantic recall [11].
3) *Augmented Generation:* Retrieved chunks are concatenated and passed to the LLM to generate an answer.

This structure has been validated across multiple domains, but its reliance on flat sequences introduces redundancy and noise. When chunks lack explicit relational encoding, models may overfit on superficial overlaps (e.g., lexical similarity) rather than extracting deeper conceptual links [7]. Furthermore, traditional RAG pipelines often perform poorly in domains with long, interdependent texts, where hierarchical relationships between concepts are essential.

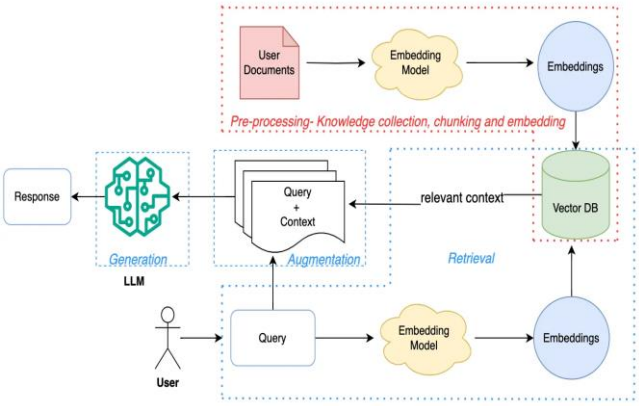

Figure 1. Illustration of RAG Architecture [30].

B. GraphRAG Structure

GraphRAG modifies the retrieval stage by introducing graph-based indexing and context construction. Instead of treating documents as isolated chunks, GraphRAG builds graphs where:

- Nodes correspond to entities, sentences, or topical communities.
- Edges capture relationships such as semantic similarity, co-reference, or hierarchy.
- Community Reports or hierarchical summaries can be generated for high-level nodes to reduce redundancy [8].

At inference time, queries are mapped onto a graph, and retrieval proceeds by exploring neighbourhoods or hierarchical paths. As illustrated in Fig. 2, query encoding, graph traversal, and subgraph selection determine the context passed to the language model, enabling structurally coherent context construction prior to generation. This results in evidence that is both structurally coherent and semantically accurate. For example, in multi-hop question answering, GraphRAG can traverse connections between entities across documents, yielding more precise retrieval contexts than top-k flat retrieval [9].

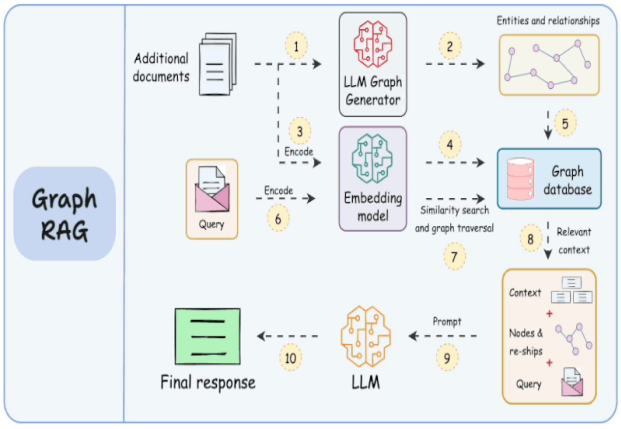

Figure 2. Illustration of GraphRAG Architecture [31].

C. Trend of Traditional RAG Research

Since its inception, Retrieval-Augmented Generation (RAG) has evolved from a hybrid of retrieval systems and neural generation into a widely adopted architecture for knowledge intensive tasks [1]. Foundational work formalised the integration of dense retrieval with sequence generation, addressing the limitations of parametric memory in LLMs by enabling access to external corpora at inference time [1][2]. The classical RAG pipeline consists of a dense retriever (e.g., Dense Passage Retrieval or vector search) that embeds queries and document chunks into a shared semantic space, followed by a generative model that conditions on the retrieved context to produce grounded outputs [10][11].

Subsequent research refined this basic model in two main ways. First, retrieval architectures were improved through hybrid sparse–dense retrievers and reranking mechanisms to enhance precision and recall, particularly in long-tail or specialised domains [10][16]. These approaches balance lexical matching with semantic similarity. Second, adaptive and feedback-driven retrieval methods emerged, where initial retrieval results guide query reformulation or iterative retrieval loops, supporting multi-step reasoning in complex QA tasks [16][17].

In parallel, work on generator–retriever integration introduced fusion mechanisms beyond simple concatenation. Cross-attention and relevance-weighted fusion modules were proposed to better align retrieved evidence with generation, improving factual grounding and reducing redundancy [7], [11]. Research also expanded toward multimodal and hierarchical retrieval settings, incorporating structured metadata and non-textual sources alongside text to address increasingly complex information demands [6][16].

Despite these advances, traditional RAG architectures remain limited by their flat retrieval paradigm. Treating document chunks as independent units restricts the modelling of inter-document relationships and multi-hop reasoning, as relational structure is not explicitly encoded in the retrieval process [4][8]. These limitations have directly motivated the development of structured retrieval paradigms such as GraphRAG and knowledge graph enhanced RAG, representing a shift toward relationally informed context construction [3][19].

D. Trend of GraphRAG Research

Graph-augmented retrieval architectures, commonly referred to as GraphRAG, represent a structural extension of traditional RAG in which retrieval is guided by explicit semantic relationships rather than flat similarity alone [3][8]. Instead of retrieving independent chunks, GraphRAG frameworks organise knowledge into graph structures, with nodes representing semantic units such as entities or sentences and edges encoding relationships including co-occurrence, hierarchy, or semantic proximity [3][13]. This structured representation enables multi-hop reasoning and the modelling of long-range dependencies that are typically inaccessible to flat retrieval systems [4][8]. GraphRAG research has concentrated on three core areas. The first is graph construction and indexing, where documents are transformed into graph representations that capture both local content and global relational structure [13][19]. Through entity extraction and relation inference, nodes can encode semantic linkages reflecting real-world knowledge hierarchies, allowing retrieval to consider paths and connections rather than isolated similarity scores [13][19]. The second area is graph-guided retrieval, in which queries traverse graph topology using neighbourhood expansion, community detection, or path scoring to retrieve structurally relevant evidence across multiple hops [4][6][22]. These methods are particularly suited to tasks requiring integrated evidence from multiple documents, such as multi-document QA and long-context summarisation [5][8].

The third area explores graph integration at the generation stage, where retrieved subgraphs inform attention mechanisms or intermediate reasoning steps, reducing hallucinations and improving logical coherence [7][8]. Query-centric graph designs, for example, introduce synthetic query nodes as abstractions between raw text and entity-level representations, improving retrieval efficiency and interpretability while reducing redundancy [22]. Variants such as PathRAG further refine this approach by optimising relational path selection to minimise noise and token overhead [23].

Overall, GraphRAG reflects a shift from similarity-based retrieval toward relational semantics, where retrieval becomes inherently structural rather than purely vector-driven [3][19]. This enables reasoning over entity relationships and information flows that span multiple documents, positioning GraphRAG as a promising architecture for complex, structured QA scenarios where flat chunk retrieval is insufficient [4][5][20].

While both traditional RAG and GraphRAG have been proposed for retrieval-augmented question answering, it remains unclear to what extent observed performance differences are attributable specifically to retrieval structures when other factors are held constant. In particular, the impact of flat versus graph-based retrieval on semantic alignment, contextual relevance, and grounding behaviour across different language models has not been systematically isolated under identical preprocessing, embedding, and evaluation conditions. This paper addresses these gaps through a controlled comparison of traditional RAG and GraphRAG pipelines using shared retrieval parameters, language model backbones, and evaluation metrics. The following sections describe the experimental methodology, present quantitative results across lexical, semantic, and grounding-based metrics, and discuss the implications of retrieval structure on performance, limitations, and future research directions.

III. METHODOLOGY

The methodology was comprised of the five steps detailed below.

A. Preprocessing and chunking

We segmented each document with Recursive Character Text Split (RCTS), with a chunk size of 1000 and overlap of 30 tokens. This method was selected to ensure contextual

continuity while maintaining manageable retrieval units. The 30-token overlap corresponds to approximately 3% of the total chunk size, which was considered sufficient to preserve cross-boundary semantic continuity between adjacent segments. Larger overlaps would increase redundancy and token overhead during retrieval and generation, potentially affecting efficiency. Given the relatively large 1000-token chunk size, a smaller proportional overlap was deemed reasonable compared to higher-percentage overlaps typically used for shorter text segments

B. Vectorization

Each document chunk was transformed into a dense vector representation using *nomic-embed-text*, a transformer-based text embedding model designed to capture semantic similarity in high-dimensional continuous space. The model maps different lengths of text inputs into fixed-size embeddings, that semantically related chunks are positioned closer together under cosine similarity, enabling effective nearest-neighbour retrieval. The embedding step is the foundation for downstream retrieval by encoding both lexical content and higher-level semantic relationships, allowing conceptually aligned statements from the document to be retrieved even when surface wording differs. Dense vectorization is important for the dataset used in this experiment, where paraphrasing and domain-specific terminology are common and exact keyword overlap is insufficient. The resulting embeddings were indexed in a vector database to support efficient similarity search during inference.

C. Preprocessing and chunking Retrieval

For the traditional RAG model, the retriever selected the top-*k* text chunks based on cosine similarity. While for the GraphRAG, the retriever performed neighbourhood exploration through document-level graph, where the nodes represented semantically related sentences, and the edges represented cosine similarity relationships.

D. Language Models and Answer Generation

To isolate the effect of retrieval structure while controlling for generative capacity, two large language models hosted via Ollama were used consistently across both RAG and GraphRAG pipelines.

The first model, LLaMA 3.2 3B, is a lightweight decoder-only transformer with approximately 3 billion parameters. It uses Grouped-Query Attention for efficient inference and has been trained on a large multilingual corpus with instruction tuning to improve reasoning and dialogue quality. While compact, it provides strong baseline performance in retrieval-augmented settings and is well suited for evaluating scenarios where retrieval quality is the primary performance bottleneck. However, its limited parameter count, and shorter effective reasoning depth may constrain performance on tasks requiring complex compositional inference [15].

The second model, Phi-4, is a reasoning-optimised decoder-only transformer developed by Microsoft, trained on high-quality curated data and synthetic textbook-style reasoning corpora. With a larger parameter count and a 16k-token context capacity, Phi-4 is better suited for structured inference and multi-hop reasoning. This makes it particularly effective in GraphRAG pipelines, where graph-structured retrieval provides richer contextual signals that can be more fully exploited by a higher-capacity model. Although computationally heavier, Phi-4 has been shown to exhibit stronger faithfulness and contextual alignment in question answering tasks [16].

For both models, retrieved context was concatenated and passed to the language model using a structured prompt template, ensuring consistent generation conditions across experimental settings. Both pipelines used nomic-embed-text for embedding and similarity computation to avoid embedding-induced bias.

E. Preprocessing and chunking Retrieval Evaluation

The generated answers were compared with the ground truth reference answers using both lexical (BLEU, ROUGE) and semantic (RAGAS) metrics. To comprehensively assess RAG performance on open-ended question answering, we adopted a multi-dimensional evaluation framework combining lexical overlap metrics, semantic similarity measures based on vector space modelling [29], and retrieval-grounding metrics [28]. This choice reflects the need to evaluate not only surface-form similarity, but also meaning preservation, factual grounding, and query alignment, which are critical in government policy QA where correct paraphrasing is often preferable to verbatim reproduction.

1) BLEU (Bilingual Evaluation Understudy): BLEU is a precision-oriented n-gram overlap metric used for automatic evaluation of machine translation [26]. It measures lexical similarity between a generated answer and a reference text. It is defined as:

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

where p_n is the modified n-gram precision, w_n are uniform weights, and BP is the brevity penalty. BLEU is included to ensure comparability with prior RAG literature but is expected to yield low scores due to the abstractive and paraphrastic nature of responses. The brevity penalty BP is defined as:

$$BP = \begin{cases} 1, & c > r \\ \exp\left(1 - \frac{r}{c}\right), & c \le r \end{cases} \quad (2)$$

where:

- c is the total length (in tokens) of the generated answer, and
- r is the total length (in tokens) of the reference answer.

2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a recall-oriented evaluation metric commonly used for summarisation and long-form text generation [27]. Unlike BLEU, which emphasises precision, ROUGE measures how much of the reference content is

covered by the generated answer. In this study, ROUGE-1, ROUGE-2, and ROUGE-L were used to capture different aspects of content overlap.

ROUGE-*N* measures *n*-gram recall and is defined as:

$$ROUGE - N = \frac{\sum_{g \in Ref} \min(Count_{gen}(g), Count_{ref}(g))}{\sum_{g \in Ref} Count_{ref}(g)} \quad (3)$$

where:

- g denotes an *n*-gram,
- G_N(Ref) is the set of all *n*-grams of length *N* appearing in the reference answer,
- $Count_{ref}$(g) is the number of times *n*-gram g appears in the reference text, and
- $Count_{gen}$(g) is the number of times *n*-gram g appears in the generated answer.

ROUGE-1 corresponds to unigram (N=1) overlap and primarily captures coverage of individual content words and key terms. ROUGE-2 corresponds to bigram (N=2) overlap and captures short phrase-level consistency and local fluency.

ROUGE-L differs fundamentally from ROUGE-1 and ROUGE-2 in that it does not rely on fixed-length *n*-grams. Instead, it measures the Longest Common Subsequence (LCS) between the generated answer and the reference text. The LCS captures the longest sequence of tokens that appear in both texts in the same order, though not necessarily contiguously. As a result, ROUGE-L is more sensitive to global sentence structure and discourse-level coherence, making it particularly suitable for evaluating longer, abstractive answers where key ideas may be reordered or paraphrased.

3) Cosine Similarity (Semantic Similarity): Cosine similarity is a standard metric in vector space information retrieval models [29]. It evaluates semantic alignment between generated answers and reference texts using embedding representations:

$$\cos(\theta) = \frac{u.v}{||u||\,||v||} \quad (4)$$

where u and v are the embedding vectors of the generated and reference texts. This metric captures meaning equivalence independent of surface form and serves as the primary indicator of semantic fidelity in abstractive dataset QA**.**

4) Faithfulness (RAGAS): Faithfulness measures whether claims in the generated answer are supported by retrieved context rather than hallucinated. Using the RAGAS framework, faithfulness is computed as the proportion of answer statements that can be grounded in retrieved evidence. This metric is critical for policy QA, where evidence-backed responses are required.

5) Context Relevance (RAGAS): Context relevance measures how directly the generated answer addresses the user query, independent of factual correctness. It quantifies semantic alignment between the query and the answer and penalises responses that are generic or tangential, providing insight into retrieval quality and query focus.

6) Metric Selection Rationale: Together, these metrics form a complementary evaluation suite:

- BLEU and ROUGE assess lexical overlap and content coverage.
- Cosine similarity captures semantic equivalence.
- Faithfulness evaluates grounding and hallucination control.
- Context relevance measures query responsiveness.

This multi-metric approach is necessary because no single metric adequately captures the performance of retrieval-augmented systems on open-ended, policy-driven QA tasks. In particular, reliance on lexical metrics alone would obscure meaningful improvements in semantic grounding and retrieval structure precisely, the aspects that GraphRAG is designed to enhance.

IV. PERFORMANCE EVALUATION

The evaluation was conducted on a policy-oriented question answering dataset derived from the dh-gen-ai-intensive-course-capstone-2025q1 corpus. The dataset consists of eight government coalition policy documents covering national reforms across governance, digital policy, economic growth, education, foreign policy, migration, and democratic stability. These documents are long-form, concept-dense, and highly abstractive, making them a challenging benchmark for retrieval-augmented generation. A total of 19 open-ended analytical questions were designed to probe factual recall, semantic understanding, multi-document integration, and policy reasoning. Two retrieval paradigms were evaluated: traditional RAG and GraphRAG. Each paradigm was tested under two language models namely LLaMA 3.2 (3B) and Phi-4.

All experiments used identical embeddings (nomic-embed-text), chunking strategy (RCTS), chunk size (1000 tokens), and overlap (30 tokens), ensuring that observed differences are attributable to retrieval structure and model behaviour rather than preprocessing variance.

To contextualise the results, it is important to note that lexical metrics such as BLEU and ROUGE typically yield modest values in open-ended, abstractive QA tasks, particularly when multiple valid phrasings exist. BLEU scores below 0.15 are common in long-form generation, while ROUGE-1 scores in the range of 0.20 - 0.30 generally indicate meaningful content overlap. In contrast, cosine similarity values above 0.65 suggest strong semantic alignment in embedding space. RAGAS-based metrics (faithfulness and context relevance) are relative measures rather than absolute quality indicators and are most informative when comparing systems under identical experimental conditions.

The results, using the various metrics, are presented in Table 1. BLEU scores are consistently low across all configurations, ranging from approximately 0.002 to 0.12, which is expected for open-ended, abstractive policy QA. BLEU is highly sensitive to exact wording and penalises paraphrasing, synonym usage, and sentence restructuring. Across both LLaMA 3B and Phi-4, traditional RAG

marginally outperforms GraphRAG on average BLEU, suggesting that flat retrieval occasionally reproduces phrasing closer to the reference text. However, this does not imply superior answer quality, as policy responses typically require abstraction and synthesis rather than verbatim reproduction. These observations are consistent with prior findings showing that BLEU correlates poorly with semantic correctness and factual adequacy in abstractive question answering and retrieval-augmented generation systems, where multiple valid phrasings exist for a given answer [12][13].

TABLE I. AVERAGE PERFORMANCE

Metric	RAG (Llama 3B)	GraphRAG (Llama 3B)	RAG (Phi4)	GraphRAG (Phi4)
BLEU	0.0278	0.0247	0.0358	0.0273
ROUGE-1	0.2128	0.2283	0.2540	0.2283
ROUGE-2	0.0715	0.0729	0.0816	0.0748
ROUGE-L	0.1586	0.1569	0.1808	0.1629
Cosine Similarity	0.6776	0.6925	0.6368	0.6847
Faithfulness (RAGAS)	0.3085	0.2866	0.2887	0.3534
Context Relevance	0.2957	0.3093	0.2426	0.3263

ROUGE metrics provide a more informative signal than BLEU for this task. Across all models and we can see that GraphRAG consistently matches or outperforms RAG on ROUGE-1 and ROUGE-2 and ROUGE-L improvements are modest but systematic. For example, in questions requiring multi-section integration (e.g., digital sovereignty, international cooperation, cybersecurity), GraphRAG demonstrates higher ROUGE-1 and ROUGE-L scores, indicating better recall of salient policy concepts and more coherent structural alignment. This suggests that GraphRAG retrieves a more complete and topically coherent subset of evidence, even when exact phrasing differs from the ground truth.

Cosine similarity emerges as the most discriminative metric for this evaluation. Across nearly all questions, we note that GraphRAG achieves higher semantic similarity than RAG and the improvement is especially pronounced with Phi-4. Using LLaMA 3B, GraphRAG improves average cosine similarity from 0.6776 to 0.6925, while using Phi-4 the improvement is even larger (0.6368 to 0.6847). These values fall well within the range expected for high-quality abstractive QA, indicating that GraphRAG better preserves the meaning of policy answers even when surface wording diverges. This result directly reflects the structural advantage of GraphRAG, which retrieves semantically linked evidence paths rather than isolated text chunks.

For the Faithfulness measure, results show a model-dependent interaction: with LLaMA 3B, traditional RAG slightly outperforms GraphRAG; with Phi-4, GraphRAG shows a substantial improvement, achieving the highest overall faithfulness score (0.3534). This indicates that GraphRAG benefits more strongly from stronger reasoning-oriented models, which are better able to exploit structured retrieval signals and relational evidence during generation.

Context relevance consistently favours GraphRAG across both models, with the largest gains observed when using Phi-4. GraphRAG achieves an average relevance score of 0.3263, compared to 0.2426 for RAG. This confirms that GraphRAG's structural retrieval mechanism helps suppress retrieval noise and maintain tighter alignment between the user query and the generated response an essential property for policy QA where tangential accuracy is insufficient.

Overall, several key patterns emerge from the results:

- Lexical metrics underestimate GraphRAG performance, particularly for abstractive answers.
- Semantic and grounding metrics consistently favour GraphRAG, especially when paired with stronger models.
- Phi-4 amplifies GraphRAG's advantages, suggesting that structured retrieval and advanced reasoning capabilities are complementary.
- Flat RAG remains competitive on surface overlap, but struggles with semantic integration, relevance control, and factual grounding.

These findings support the hypothesis that GraphRAG represents a structural evolution of RAG, particularly well-suited for complex, multi-document, policy-driven question answering.

A. Limitations

The primary limitations observed in this study arise from the interaction between model capacity, retrieval structure, domain specificity, and evaluation methodology. Both traditional RAG and GraphRAG operate over policy documents that are inherently abstractive, normative, and semantically dense, which substantially reduces lexical overlap with reference answers and leads to systematically low BLEU and moderate ROUGE scores despite semantically correct outputs. Traditional RAG is further constrained by its flat retrieval paradigm, which treats document chunks as independent units and therefore fails to model inter-document relationships or multi-hop reasoning paths that are common in coalition policy questions. GraphRAG mitigates this issue through structured retrieval but introduces its own limitations, particularly sensitivity to graph construction quality, entity linking noise, and traversal depth selection, which can propagate weakly relevant nodes into the retrieved context and slightly depress faithfulness scores. Across both approaches, the use of general-purpose embedding models and relatively small language models limits fine-grained representation of policy-specific terminology and long-range reasoning, while reliance on single-reference lexical metrics underestimates answer quality by penalising valid paraphrasing, synthesis, and

abstraction that are essential for policy-oriented question answering.

V. CONCLUSION AND FUTURE WORK

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

While both retrieval-augmented approaches exhibit limitations under traditional lexical evaluation metrics, the results demonstrate that GraphRAG provides a structurally superior and more future-proof framework for complex, policy-oriented question answering. Traditional RAG remains competitive in constrained scenarios where surface-level retrieval suffices, but its flat architecture limits multi-hop reasoning and relational synthesis across documents. GraphRAG consistently achieves stronger semantic alignment and query relevance, particularly when paired with more capable language models, and its remaining weaknesses are largely attributable to implementation maturity rather than conceptual design. Overall, the evidence supports GraphRAG as the more effective and scalable architecture for high-stakes, knowledge-intensive applications where correctness, grounding, and interpretability outweigh surface-form similarity.

Future work will focus on improving graph construction precision, node weighting, and dynamic subgraph selection to reduce retrieval noise, as well as exploring hybrid retrieval strategies and adaptive chunking to improve evidence coverage. Additional gains may be achieved through embedding fine-tuning, planning-aware and citation-constrained generation, and the use of models with stronger long-context reasoning capabilities. Finally, future evaluation should prioritise semantic and grounding-based metrics with multiple reference answers to better reflect the abstractive nature of policy QA, as low BLEU and moderate ROUGE scores primarily reflect metric mismatch rather than model failure.

ACKNOWLEDGMENT

This research is funded by UKRI (UK Research and Innovation) under the Hartree National Centre for Digital Innovation Programme (Hartree Hub Northern Ireland - Home.)

REFERENCES

[1] P. Lewis, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[2] G. Izacard and E. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering,” arXiv preprint arXiv:2007.01282, 2020.

[3] M. Yasunaga, X. Ren, B. Bosselut, P. Liang, and J. Leskovec, “GraphRAG: Retrieval-Augmented Generation with Graphs,” arXiv preprint arXiv:2501.00309, 2025.

[4] Y. Zhao, Z. Wang, H. Chen, and J. Li, “When to Use Graphs in Retrieval-Augmented Generation: A Study on Complex Question Answering,” arXiv preprint arXiv:2506.05690, 2025.

[5] D. Zerva, K. Kapanipathi, V. Karpukhin, and S. Riedel, “GraphRAG-Bench: A Benchmark for Graph-Based Retrieval-Augmented Generation,” arXiv preprint arXiv:2506.02404, 2025.

[6] J. Lin, Z. Wang, Y. Li, and X. Ren, “From Local to Global: A Graph-RAG Approach to Query-Focused Summarization,” arXiv preprint arXiv:2404.16130, 2024.

[7] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, “Retrieval Augmentation Reduces Hallucination in Conversation,” in Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 3784–3803, 2021.

[8] H. Zhang, Y. Li, J. Gao, and Z. Liu, “RAG vs. GraphRAG: A Systematic Evaluation and Key Insights,” arXiv preprint arXiv:2502.11371, 2025.

[9] Deep-PolyU Research Group, “Awesome GraphRAG,” GitHub repository. [Online]. Available: https://github.com/Deep-PolyU/Awesome-GraphRAG, 2025.

[10] V. Karpukhin, et al., “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, 2020.

[11] L. Gao, X. Ma, J. Lin, and Z. Liu, “Rethink Training of Retrieval-Augmented Generation for Open-Domain Question Answering,” in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2061–2077, 2022.

[12] R. Zhong, M. Yasunaga, and X. Ren, “Extractive Retrieval-Augmented Generation for Policy and Regulatory Documents,” arXiv preprint, 2023.

[13] M. Yasunaga, H. Chen, Y. Ren, and P. Liang, “Graph-Based Retrieval and Reasoning for Open-Domain Question Answering,” in Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 163–175, 2021.

[14] Meta AI, “LLaMA 3.2-3B,” Hugging Face Model Repository. [Online]. Available: https://huggingface.co/meta-llama, 2024.

[15] Microsoft, “Introducing Phi-4: Microsoft’s Newest Small Language Model Specializing in Complex Reasoning,” Microsoft Community Hub, 2024.

[16] S. Gupta, R. Ranjan, and S. N. Singh, “A Comprehensive Survey of Retrieval-Augmented Generation: Evolution, Current Landscape, and Future Directions,” arXiv preprint, 2024.

[17] A. Gan, et al., “Retrieval-Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey,” arXiv preprint, 2025.

[18] Y. Huang and J. Huang, “A Survey on Retrieval-Augmented Text Generation for Large Language Models,” arXiv preprint, 2024.

[19] B. Peng, Y. Zhu, Y. Liu, X. Bo, and H. Shi, “Graph Retrieval-Augmented Generation: A Survey,” arXiv preprint, 2024.

[20] Q. Zhang, et al., “A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models,” arXiv preprint, 2025.

[21] QualityPoint Technologies, “The Evolution of Retrieval-Augmented Generation,” QualityPoint Technologies Blog, 2025.

[22] Emergent Mind Research, “Graph-Centric RAG Frameworks and Query-Centric Design Studies,” Emergent Mind Overview, 2025.

[23] Research Community, “PathRAG: Path-Focused Graph Retrieval-Augmented Generation Innovations,” Community Research Summary, 2025.

[24] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, 2006, pp. 249–256.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74 - 81.

[28] S. Es, S. James, A. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," arXiv:2309.15217, 2023.[retrieved: Feb. 2026].

[29] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[30] Mindful Matrix, "Building LLM application using RAG" Substack, [Online]. Available: https://mindfulmatrix.substack.com/p/build-a-simple-llm-application-with, 2024.

[31] D. vonThenen, "Beyond Vectors – Knowledge Graphs & RAG Using Gradient," DigitalOcean Community, [Online]. Available: https://www.digitalocean.com/community/tutorials/beyond-vectors-knowledge-graphs-and-rag, 2025.