Science and Technology Facilities Council

Hartree Centre | Northern Ireland Hub

Comparative Evaluation of RAG and GraphRAG

Open-Ended Question Answering in Policy Documents

Jadesola Osinowo, Abiodun Adebayo, Sonya Coleman,
Dermot Kerr, Justin Quinn

Hartree Centre Northern Ireland Hub

SCEIS, Ulster University

*Slides prepared and presented by **Abiodun Adebayo***

Ulster University

Jade is a Research Associate in Data Science at the Hartree NI Hub, where she brings expertise across several aspects of AI, delivering specialized assists and AI solutions to industry partners.

She graduated with a BSc in Software Engineering from the University of Leicester and an MSc in Artificial Intelligence and Data Analytics from Loughborough University.

The Challenge: Finding the Right Answer

Complex Queries

Multi-hop reasoning across documents

Hidden Connections

Relationships lost in flat text

Dense Content

Legal, policy, scientific domains

Hallucinations

AI making up facts without evidence

💡 *Traditional RAG treats documents as flat sequences — missing deeper semantic relationships*

Retrieval-Augmented Generation (RAG)

Combining AI with External Knowledge

Retrieve

Search external documents for relevant information

Augment

Add retrieved context to the AI's input

Generate

AI produces factually grounded answers

✓ Key Benefits

Reduces hallucinations • Improves accuracy • Provides evidence-based answers • Current information access

Traditional RAG: The Limitations

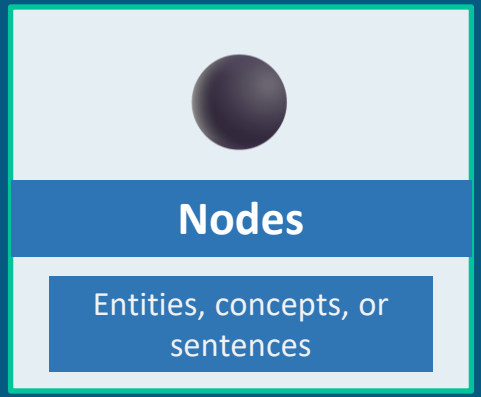
How It Works:

The Problems:

- ✗ No relationship tracking between chunks
- ✗ Struggles with multi-hop reasoning
- ✗ Misses hierarchical connections
- ✗ Can't follow entity relationships
- ✗ Relies on superficial word matching

Introducing GraphRAG

A Graph-Based Approach to Retrieval



Instead of isolated chunks, GraphRAG builds a knowledge network

How GraphRAG Works

1. **Build Graph**
   Extract entities & relationships
2. **Index Structure**
   Create searchable knowledge network
3. **Query Mapping**
   Map question to graph nodes
4. **Path Traversal**
   Follow connections to gather evidence

Advantages

- ✓ Captures relationships explicitly
- ✓ Enables multi-hop reasoning
- ✓ Reduces retrieval noise
- ✓ More coherent context
- ✓ Better semantic alignment

Research Methodology

Dataset

📄 **Source:** German coalition policy documents

📊 **Content:** 8 policy documents covering governance, digital policy, economy, education

❓ **Questions:** 19 open-ended analytical questions

🎯 **Challenge:** Long-form, concept-dense, abstractive content

Models & Configuration

🤖 **LLaMA 3.2 3B**
Lightweight, 3B parameters

🤖 **Phi-4**
Reasoning-optimized, 16k context

⚙️ **Processing:**
- 1000-token chunks, 30-token overlap
- nomic-embed-text embeddings
- Identical preprocessing for fair comparison

How We Measured Performance

Multi-Dimensional Evaluation Framework

Lexical Overlap

BLEU: Word precision

ROUGE: Content coverage

Semantic Similarity

Cosine Similarity:

Meaning equivalence

Grounding (RAGAS)

Faithfulness: Evidence-backed

Context Relevance: Query focus

Why multiple metrics? No single measure captures all aspects of answer quality in policy QA

Key Findings

Semantic Similarity

GraphRAG

0.685

vs

0.637

+7.5%

(Phi-4)

Faithfulness

GraphRAG

0.353

vs

0.289

+22%

(Phi-4)

Context Relevance

GraphRAG

0.326

vs

0.243

+34%

(Phi-4)

The Model Matters

GraphRAG + Phi-4 = Best Performance

Metric	RAG (LLaMA 3B)	GraphRAG (LLaMA 3B)	RAG (Phi-4)	GraphRAG (Phi-4)
ROUGE-1	0.213	0.228	0.254	**0.228**
Cosine Sim	0.678	0.693	0.637	**0.685**
Faithfulness	0.309	0.287	0.289	**0.353**
Context Rel.	0.296	0.309	0.243	**0.326**

🎯 **Key Insight: Phi-4's stronger reasoning capabilities amplified GraphRAG's structural advantages**

Understanding Low Lexical Scores

Why BLEU and ROUGE Scores Were Low

What These Metrics Measure

BLEU: Exact word matching

ROUGE: Word and phrase overlap

Both penalize:
- Paraphrasing
- Using synonyms
- Sentence restructuring
- Abstraction & synthesis

The Reality

Policy QA requires:
- Synthesis across documents
- Paraphrasing complex concepts
- Abstractive reasoning

Multiple valid phrasings exist for the same answer

Low lexical scores ≠ poor quality
They reflect the abstractive nature of the task

When Should You Use GraphRAG?

GraphRAG Shines

- ✓ Multi-document reasoning
- ✓ Policy and legal analysis
- ✓ Scientific literature review
- ✓ Complex knowledge domains
- ✓ Entity relationship queries
- ✓ Multi-hop question answering

Traditional RAG Works

- Simple fact lookup
- Single-document queries
- Quick factoid questions
- Surface-level retrieval
- When speed is critical
- Limited computational resources

Limitations & Future Directions

Current Limitations

- ⚠️ Graph construction quality affects performance
- ⚠️ Sensitive to entity linking accuracy
- ⚠️ Computational overhead vs traditional RAG
- ⚠️ Limited by small LLM capacity (3B parameters)

Future Research

- → Improved graph construction and node weighting
- → Hybrid retrieval strategies (combining flat + graph)
- → Larger models with better reasoning capabilities
- → Better evaluation metrics for abstractive tasks

Science and Technology Facilities Council

Hartree Centre | Northern Ireland Hub

Conclusion

1. GraphRAG provides superior semantic alignment and grounding
2. Especially effective for complex, policy-oriented QA tasks
3. Performance amplified when paired with stronger LLMs (Phi-4)
4. Future-proof architecture for knowledge-intensive applications

GraphRAG represents a structural evolution in retrieval-augmented generation

Thank You

Funded by UKRI - Hartree National Centre for Digital Innovation

Ulster University