A Knowledge Graph-Based RAG Approach for Question Answering

A case study on EmPULIA Regulations

Mattia Curri
Dipartimento di Informatica
Università degli Studi di Bari Aldo Moro, Italy

July 14, 2025

The Challenge with LLMs in Specialized Domains

Critical Limitations in Public Administration

Large Language Models face significant challenges when applied to specialized fields:

- Hallucinations: Generating plausible but factually incorrect information
- Outdated Knowledge: Lack of access to current, domain-specific information
- Poor Verifiability: Difficulty tracing sources - crucial for accountability
- Opacity: Internal decision-making processes make verification challenging

Why This Matters

In public administration, accessing accurate and up-to-date information is **critical**.

Our Solution: Knowledge Graph-Enhanced RAG

Knowledge Graphs (KGs)

- Represent structured, interconnected knowledge
- Excel at complex relationship representation
- Enable structured, connected knowledge representation

Retrieval-Augmented Generation (RAG)

- Grounds LLM responses in retrieved, relevant data
- Mitigates hallucinations and improves accuracy
- Enhances LLMs with factual, domain-specific information

Goal: Create an accurate, verifiable, and context-aware QA system for **EmPULIA** e-procurement platform regulations.

Related Work: The Evolution of RAG

Standard RAG Limitations

Standard RAG approaches have limitations:

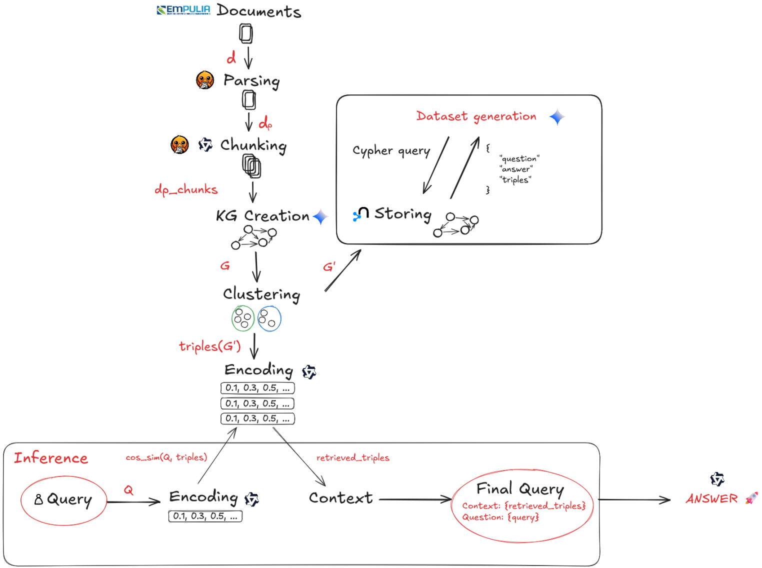
- Struggle with complex queries requiring relationship understanding
- Limited ability to navigate interconnected information
- Difficulty with multi-hop reasoning tasks
- Insufficient for intricate regulatory document analysis

Knowledge Graph Integration

Recent advances integrate KGs with RAG:

- **GraphRAG**: Leverages graph-based index for relationship navigation
- Better handling of complex questions requiring relationship understanding

High-Level Pipeline Overview



Document Processing Pipeline

Document Collection & Parsing

- Official regulatory texts from `EmPULIA` platform
- Parse PDFs using `Docling` with `HybridChunker`
- The `HybridChunker` ensures semantically coherent segments while maintaining tokenization awareness for better downstream processing.
- Uses `Qwen3-Embedding-0.6B` tokenizer for consistent boundaries
- Aggregate every three consecutive chunks for optimal granularity

Knowledge Graph Construction: Technical Details

Triple Extraction Process

- `Gemini-2.5-Flash` with structured output
- **Parameters:** `temperature=0`, `thinking_budget=0`
- Extract format: `(entity1, relation, entity2, source)`
- Merge all triples across chunks into unified KG

Graph Refinement & Clustering

- Cluster relation labels using `DBSCAN`
- **Parameters:** $\epsilon = 0.05$ for high similarity clustering
- Results in more cohesive and compact graph structure

Triple Encoding and Storage Strategy

Embedding Strategy

- Embed triples using `Qwen3-Embedding-0.6B`
- **Format:** `entity1 relation entity2: source`
- Alternative: embed only core triple `entity1 relation entity2`

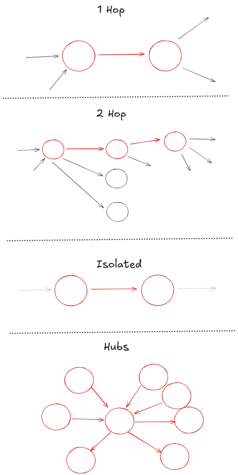
Retrieval Configuration

- **Search parameters:** `cosine_threshold=0.6`, `top_k=546`
- High k-value (size of largest connected component) tests model's filtering ability
- Two embedding strategies evaluated: with/without source information

EmPULIA-QA Dataset Construction

Systematic Difficulty Levels

- **Isolated Nodes**: Simple fact retrieval from single entities (8 questions)
- **Single Hop**: Questions requiring one relationship traversal (8 questions)
- **Two Hops**: Multi-step reasoning across two relationships (8 questions)
- **Hubs**: Queries about highly connected entities (6 questions)
- **Out-of-Domain**: Irrelevant questions (6 questions)



Answer Generation: Technical Implementation

Qwen3-4B Configuration

- **Core parameters:** `temperature=0`, `contextlength=8192`
- **Sampling:** `top_p=0.95`, `top_k=20` (Ollama default)
- **Quantization:** `Q4_K_M` (Ollama base quantization)

Structured Response Format

Each generated response contains three components:

- **Answer**: Direct reply to the user's query, grounded in context
- **Analysis**: Explanation of reasoning process and context usage
- **Sources**: Specific KG triples cited, ensuring full verifiability

Comprehensive Evaluation Metrics

Retrieval Component

- **Context Precision**:

$$CP = \frac{\sum(Precision@k)}{\text{No. relevant items in top K}}$$

- **Context Recall**:

$$CR = \frac{\sum(Recall@k)}{\text{No. relevant items in top K}}$$

Generation Component

- **Context Faithfulness**: Can generated answer be inferred from context?
- **Analysis Faithfulness**: Is reasoning explanation supported by context AND answer?
- **Answer Accuracy**: agreement between a model's response and the ground truth. LLM jury evaluation with perspective swapping (0-4 scale, then normalized)

Results - Retrieval Component

EmPULIA-QA						
Category	**Precision**			**Recall**		
	Source	w/out Source	Δ	Source	w/out Source	Δ
Single Hop	**0.539**	0.394	↓ 0.145	0.875	0.875	– 0.000
Two Hop	**0.659**	0.520	↓ 0.139	**1.000**	0.750	↓ 0.250
Isolated	0.567	**0.660**	↑ 0.093	**1.000**	**1.000**	– 0.000
Hubs	**0.463**	0.415	↓ 0.048	**0.108**	0.061	↓ 0.047
Out of Domain	1.000	1.000	– 0.000	1.000	1.000	– 0.000

Table: Retrieval Evaluation on `EmPULIA-QA` dataset. Precision and recall for each category, comparing *source* vs *nosource*. Best values in bold. Arrows indicate variation (nosource - source).

Results - Generation Component

EmPULIA-QA									
Category	**Context Faithfulness**			**Analysis Faithfulness**			**Answer Accuracy**		
	Source	w/out Source	Δ	Source	w/out Source	Δ	Source	w/out Source	Δ
Single Hop	0.288	**0.438**	↑ 0.150	0.625	**0.740**	↑ 0.115	0.250	0.250	- 0.000
Two Hop	**0.769**	0.583	↓ 0.186	**0.924**	0.913	↓ 0.011	0.375	**0.406**	↑ 0.031
Isolated	**0.627**	0.552	↓ 0.075	0.862	**0.958**	↑ 0.096	0.531	**0.594**	↑ 0.063
Hubs	0.565	**0.644**	↑ 0.079	0.844	**0.921**	↑ 0.077	0.333	**0.500**	↑ 0.167
Out of Domain	1.000	1.000	− 0.000	1.000	1.000	- 0.000	1.000	1.000	- 0.000

Table: Generation Evaluation on `EmPULIA-QA` dataset. Faithfulness and answer accuracy per category. Best source values in bold. Arrows indicate variation (nosource - source).

Source

Answer: L'ammontare a base d'asta che definisce gli appalti di sola esecuzione è di 100 milioni di euro, come stabilito nell'Articolo 129, comma 3 del dPR 5 ottobre 2010.
Analysis: Il contesto specifica che gli appalti di sola esecuzione hanno una base d'asta superiore a 100 milioni di euro, con riferimento all'Articolo 129, comma 3 del dPR 5 ottobre 2010. Questo articolo è citato come fonte per la definizione dell'importo minimo.
Sources:

- Fonte: dPR 5 ottobre 2010_207_agg_DM_infrastrutture_24apr2014, Art. 129, comma 3: «...di ammontare a base d'asta superiore a 100 milioni di euro.»

No Source

Answer: Il limite di base d'asta per gli appalti di sola esecuzione è di 100 milioni di euro, come stabilito nell'articolo 129, comma 3 del dPR 5 ottobre 2010. Questo è confermato dal riferimento al testo dPR 5 ottobre 2010_207_agg_DM_infrastrutture_24apr2014, art. 129, comma 3.

Analysis: La domanda chiede il limite di base d'asta. Il contesto specifica che gli appalti di sola esecuzione hanno una base d'asta superiore a 100 milioni di euro, riferendosi all'articolo 129, comma 3 del dPR 5 ottobre 2010_207_agg_DM_infrastrutture_24apr2014. Questo articolo è citato come fonte per il limite. Non ci sono altre informazioni aggiuntive nel contesto.

Sources:

- Fonte: dPR 5 ottobre 2010_207_agg_DM_infrastrutture_24apr2014, Art. 129, comma 3: «...di ammontare a base d'asta superiore a 100 milioni di euro.»

Question

Question: Quali sono le disposizioni relative ai materiali non accettati e ai magazzini secondo l'Art. 109, comma 6 del D.Lgs. 50/2016?

Answer Accuracy Evaluation

- Score 2: La risposta generata è parzialmente allineata alla ground truth, ma omette una delle due disposizioni chiave relative ai magazzini a disposizione della Stazione Appaltante.
- Score 0: La risposta generata non corrisponde alla ground truth. La ground truth indica solo la rimozione dei materiali non accettati, **mentre la risposta generata aggiunge una seconda disposizione non presente nella ground truth.**

Generated Response

Answer: Il comma 6 dell'art. 109 del D.Lgs. 50/2016 prevede che l'appaltatore deve rimuovere dai magazzini e dai cantieri i materiali non accettati dal direttore dei lavori. **Inoltre, deve mettere i magazzini a disposizione della stazione appaltante.**

Analysis: Il comma 6 dell'art. 109 stabilisce due obblighi principali per l'appaltatore:

1. Rimuovere i materiali non accettati dai magazzini e dai cantieri.
2. Mettere i magazzini a disposizione della stazione appaltante.

Sources:

... L'appaltatore deve rimuovere dai magazzini e dai cantieri i materiali non accettati dal direttore dei lavori.

... Deve mettere i magazzini a disposizione della stazione appaltante.

Some Insights

1. Data Preprocessing (Still) Matters

- Throwing LLMs at problems doesn't guarantee good results
- Quality of input data and preprocessing steps are **crucial**
- Careful document parsing and chunking significantly impact performance

2. Knowledge Graphs Are Powerful (If Carefully Crafted)

- Well-designed KGs significantly enhance retrieval and generation
- Structured, domain-specific context provides substantial benefits

3. Small Language Models Are Not Enough (Yet)

- Even with proper context, smaller models struggle with effective utilization
- Context filtering and relevance assessment remain challenging
- Needle in the Haystack problem persists

Current Limitations

Limitations

- **Graph topology**: Current embedding treats triples independently
- **Source validation**: Citation generation not rigorously evaluated
- **Dataset coverage**: Some regulatory facts may lie outside current scope
- Limited to EmPULIA domain - generalization needs exploration

Future Work

- **Prompt refinement**: Leverage model's instruction-following capabilities
- **Graph enhancement**: Utilize Neo4j's node/relation property capabilities
- **User feedback**

Thank You for the softmax $\left(\frac{QK^T}{\sqrt{d_k}}\right) V$

🚀