PHAROS AI Factory · AI4Language · 07 July 2026

PHAROS
THE GREEK AI FACTORY

Retrieval-Augmented Generation (RAG) End-to-End: Architecture, Retrieval, Generation and Evaluation

https://events.grnet.gr/event/213/

George Drosatos · Sotirios Gyftopoulos
Institute for Language and Speech Processing, Athena Research Center
Emails: gdrosato@athenarc.gr, sotiris.gyftopoulos@athenarc.gr

ATHENA
Research & Innovation
Information Technologies

Webinar path and schedule

11:00–11:15	Why RAG? Hallucinations, knowledge cut-offs and grounded generation	**13:05–13:30**	Retrieval strategies, HyDE, RRF and reranking
11:15–11:40	RAG architectures: Naïve, Advanced and Modular RAG	**13:30–13:55**	Hands-on 3: semantic and hybrid retrieval
11:40–12:10	Hands-on 1: dataset loading and first RAG overview	**13:55–14:20**	Generation: context, grounding, citations
12:10–12:35	Ingestion, chunking, metadata and embeddings	**14:20–14:40**	Hands-on 4: end-to-end RAG chain
12:35–12:50	Hands-on 2: chunking and vector indexing	**14:40–15:00**	RAG evaluation: quality, faithfulness and frameworks
12:50–13:05	Break	**15:00–15:15**	Deployment considerations and wrap-up

Course orientation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

What participants will be able to do

- Explain the architecture and main paradigms of RAG
- Design a pipeline from document preparation to answer generation
- Make informed choices about chunking, embeddings, retrieval and prompting
- Evaluate retrieval quality, answer quality and end-to-end behaviour

Practical focus

A Greek-language public-service assistant will be used as the running example

Outcome

Participants leave with a working Colab-based RAG pipeline and a mental model for deployment

Course orientation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

11:00–11:15 · RAG motivation

PHAROS
THE GREEK AI FACTORY

Why RAG?

Hallucinations, knowledge cut-offs and grounded generation

Why RAG exists

Static knowledge
A model cannot know every new policy, regulation or internal document after training

Hallucinations
Fluent answers can be fabricated when evidence is missing or ambiguous

Domain mismatch
Specialised terminology and Greek public-sector language need approved sources

RAG changes the task from “generate from memory” to “answer from retrieved evidence”

RAG motivation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

RAG in one sentence

Retrieval-Augmented Generation retrieves external knowledge at inference time and uses it to condition the LLM's response

- The external source can be a document collection, database, knowledge graph or hybrid index
- The generated answer should be constrained by the retrieved evidence and expose source attribution

RAG motivation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Where RAG provides value

Public services

Procedures, eligibility criteria, forms, calls and administrative answers

Enterprise knowledge

Internal policies, support articles, product documentation and manuals

Regulated domains

Legal, healthcare, finance and other evidence-sensitive contexts

- The common requirement is not just a plausible answer, but a verifiable answer

RAG motivation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

11:15–11:40 · RAG Theory

PHAROS
THE GREEK AI FACTORY

RAG architectures

Naïve, Advanced and Modular RAG

The end-to-end RAG pipeline

- **Indexing phase:** prepare the knowledge base before users ask questions
- **Query-time phase:** retrieve relevant evidence, construct context and generate the response
- **Evaluation phase:** measure whether retrieval and generation behaved as expected

RAG architecture

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Naïve RAG: the baseline architecture

1. Ingestion / indexing
Clean documents, split them into chunks, compute embeddings and store them

2. Retrieval
Encode the query and select the most relevant chunks, often with Top-K similarity

3. Generation
Inject retrieved context into the prompt and generate a grounded answer

- Good baseline for understanding RAG, but sensitive to chunking, retrieval quality, context noise and prompt design

RAG architecture

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Advanced RAG: improve before, during and after retrieval

Pre-retrieval

- Structure-aware chunking
- Metadata enrichment
- Query rewriting / expansion
- HyDE-style query transformation

Retrieval

- Dense semantic search
- Sparse keyword search
- Hybrid retrieval
- Adaptive Top-K selection

Post-retrieval

- Fusion
- Reranking
- Context compression
- Evidence filtering

- Advanced RAG is about reducing retrieval noise and increasing evidence coverage before the LLM sees the context

RAG architecture

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Modular RAG: configurable retrieval workflows

- Routing decides which knowledge source or retriever to use
- Query rewriting and decomposition help with vague or multi-part questions
- Verification modules detect unsupported claims and trigger re-retrieval or clarification

The key idea: the pipeline becomes adaptive rather than fixed

RAG architecture

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

11:40–12:10 · Tutorial

PHAROS
THE GREEK AI FACTORY

Hands-on 1

Dataset loading, document preparation and first RAG overview

Running example: Greek public-service assistant

User need

Ask questions in natural Greek about public-service procedures or eligibility

Knowledge source

Structured and unstructured documents, Q&A records and administrative text

System response

Grounded answer with source attribution, caveats and clear wording

- This scenario is useful because it combines Greek language, domain terminology, evidence sensitivity and a realistic user interface

Hands-on 1

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Notebook structure: from documents to answers

- Each notebook section corresponds to one design decision in the RAG pipeline
- Intermediate outputs are inspected so participants can see how data changes at each stage
- The final result is a runnable pipeline rather than isolated code fragments

Hands-on 1

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

12:10–12:35 · Methodology

PHAROS
THE GREEK AI FACTORY

Ingestion, chunking, metadata and embeddings

Preparing knowledge so it can be retrieved

Data preparation: preserve meaning before indexing

- Extract text from PDFs, HTML, Markdown or structured files
- Remove boilerplate, duplicated fragments and noisy artefacts
- Preserve headings, sections, tables, article numbers and source IDs
- Normalise encoding, whitespace and domain-specific notation

Design principle

The parser should preserve the structure that users will later ask about

Public-sector example

Calls, articles, tables and eligibility criteria must remain traceable

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Chunking: balancing coherence and retrieval precision

Fixed-size
Simple baseline; may split meaning across boundaries

Overlapping
Preserves context across neighbouring chunks

Semantic
Splits by meaning, paragraph or sentence coherence

Structure-aware
Uses sections, articles, headings or tables

Chunk size is a retrieval decision, not just a preprocessing detail

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Metadata makes retrieval controllable

- **Useful metadata fields**
  - Document title and source URL
  - Section, article, paragraph or table ID
  - Publication date and authority
  - Language, document type and topic
  - Chunk position and parent section

Filtering

Search only in the relevant subset of the corpus

Traceability

Citations can point to the exact passage, not just the document

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Embeddings and vector indexing

- Embeddings place semantically similar chunks close to each other in vector space
- Multilingual or Greek-aware embeddings are important for Greek-language applications
- Index choice affects latency, scaling, filtering and maintainability

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

12:35–12:50 · Tutorial

PHAROS
THE GREEK AI FACTORY

Hands-on 2

Chunking, metadata and vector indexing

Hands-on 2: implementation checkpoints

- Inspect raw documents and identify logical boundaries
- Create chunks with parent metadata and source identifiers
- Generate embeddings for each chunk
- Store chunks and vectors in a local index
- Run a first similarity search and inspect the returned evidence

Checkpoint

The notebook should make it clear which text was indexed and why it is retrievable

Hands-on 2

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

12:50–13:05 · Break

PHAROS
THE GREEK AI FACTORY

15 minutes short break

13:05–13:30 · Methodology

PHAROS
THE GREEK AI FACTORY

Retrieval strategies

Semantic, lexical, hybrid retrieval, HyDE, RRF and reranking

Retrieval modalities in RAG

Dense retrieval
Embeddings capture semantic similarity and paraphrases

Sparse retrieval
Keyword methods capture exact terms, IDs and rare phrases

Structured retrieval
Metadata, tables, filters and graph relations add constraints

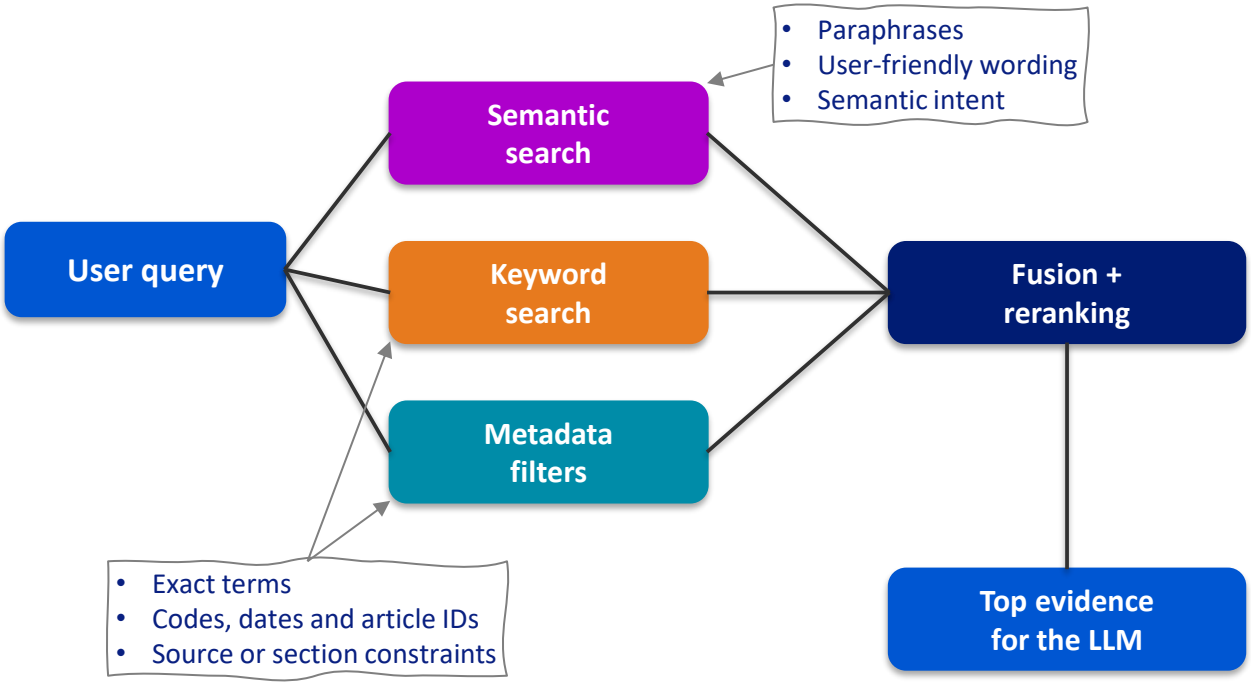
Hybrid
Combines complementary signals

- Production RAG often combines semantic recall with lexical precision and metadata control

G. Drosatos & S. Gyftopoulos, ILSP / Athena RC

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Hybrid retrieval: combine complementary signals



Retrieval

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Query enhancement before retrieval

Query rewriting
Rephrase a vague or conversational query into a search-friendly form

Query expansion
Add synonyms, domain terms or related concepts to improve recall

HyDE
Generate a hypothetical answer-like document and embed it for search

- The goal is not to change the user's intent, but to retrieve evidence that matches it

Retrieval

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Fusion and reranking: make Top-K count

- Fusion combines results from multiple retrievers without losing complementary evidence
- Reranking scores query–chunk pairs more carefully than vector similarity alone
- Top-K is both a quality decision and a context-window decision

Retrieval

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Retrieval tuning levers

Top-K
More evidence, but more noise and cost

Thresholds
Filter low-confidence results before generation

Metadata filters
Restrict search by source, date, section or authority

Reranker
Improve ordering of candidate chunks

Chunking
Change the unit of evidence being retrieved

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

13:30–13:55 · Tutorial

PHAROS
THE GREEK AI FACTORY

Hands-on 3

Semantic search, BM25/hybrid retrieval and reranking

Hands-on 3: compare retrieval strategies

- Run semantic search with representative questions
- Run keyword or hybrid search for exact administrative terms
- Inspect false positives and missing evidence
- Apply reranking and compare the Top-K context
- Record which strategy works best for each query type

Practical rule

Do not trust retrieval scores blindly. Always inspect the returned evidence

Hands-on 3

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

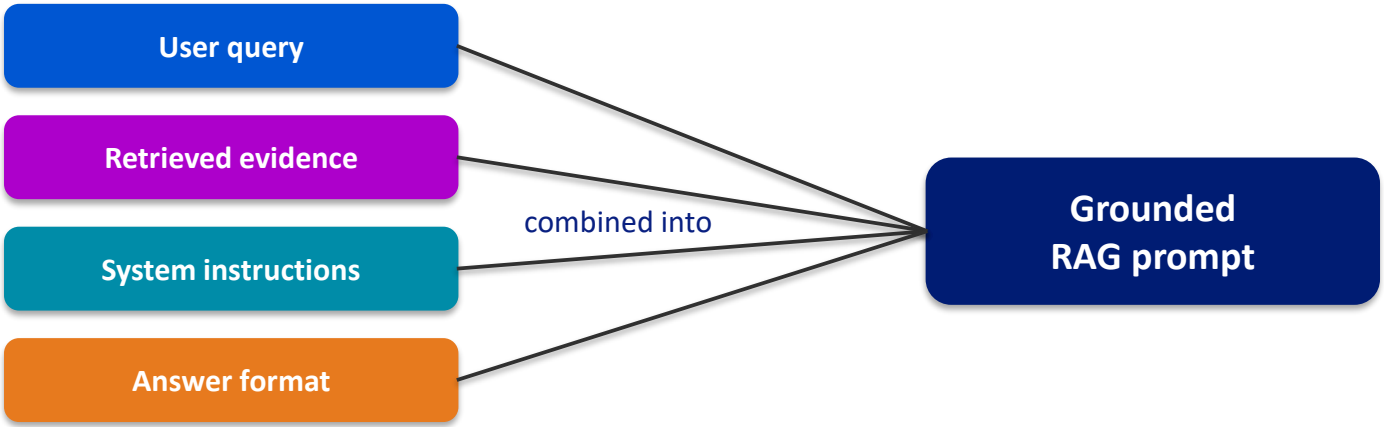
13:55–14:20 · Methodology

PHAROS
THE GREEK AI FACTORY

Generation in RAG

Context construction, grounding, citations and answer constraints

Context construction controls the answer



- The prompt should specify how to use evidence, how to cite sources and what to do when evidence is insufficient

Generation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Grounding and source attribution

Grounded answer

Claims should be supported by retrieved context, not by unsupported model knowledge

Citation granularity

Cite exact passages, sections, articles or source IDs when possible

Transparency

Make the evidence inspectable and separate facts from caveats

- In evidence-sensitive domains, a correct answer without traceability is still incomplete

Generation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

When evidence is insufficient

- State that the available sources do not contain enough information
- Ask a clarification question if the user intent is ambiguous
- Avoid unsupported legal, administrative or procedural claims
- Provide only the evidence-backed part of the answer
- Log unresolved questions for dataset improvement

Reliability pattern

It is better to say “not enough evidence” than to generate an unsupported answer

Generation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

14:20–14:40 · Tutorial

PHAROS
THE GREEK AI FACTORY

Hands-on 4

End-to-end RAG answer generation with source attribution

Hands-on 4: build the RAG chain

- Inspect the final prompt to understand what the LLM actually receives
- Compare answers with and without retrieved context
- Check whether the cited sources truly support the answer
- Prepare the chain for a simple user-facing interface

Hands-on 4

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

14:40–15:00 · Evaluation

PHAROS
THE GREEK AI FACTORY

RAG evaluation

Retrieval quality, faithfulness, groundedness and evaluation frameworks

Evaluation must decompose the pipeline

- A wrong answer can come from missing evidence, noisy context, a weak grounding prompt or hallucinated generation
- Good evaluation localises the failure and links metrics to concrete system changes

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Retrieval evaluation: what did we retrieve?

Precision@k
How clean are the top-k chunks?

Recall@k
Did we retrieve the necessary evidence?

MRR
How early is the first useful chunk?

NDCG
Are stronger evidence chunks ranked higher?

- Retrieval metrics decide what evidence the LLM is allowed to see

Evaluation

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Answer evaluation: is the response supported?

Correctness
Is the answer true in the target domain?

Faithfulness
Are the claims supported by the retrieved context?

Groundedness
Does the answer avoid unsupported external claims?

Citation quality
Are sources exact and relevant?

Completeness
Does it cover what the user asked?

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Evaluation frameworks: RAGAS and DeepEval

RAGAS

Automated RAG evaluation with metrics such as faithfulness, answer relevance, context precision and context recall. Useful for experiments and regression comparisons

DeepEval

LLM unit-testing style framework with threshold-based checks, custom metrics, RAG metrics and CI/CD-oriented workflows

- Frameworks compute scores, but the engineering value is the diagnosis behind the score

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

From metrics to failure diagnosis

Wrong answer, no evidence	**Low Recall@k**	Improve query / chunking
Noisy context	**Low Precision@k**	Add filters / reranking
Unsupported claim	**Low faithfulness**	Strengthen grounding prompt
Bad citation	**Low citation quality**	Use passage-level source IDs

- The best evaluation produces an improvement decision, not only a dashboard score

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

15:00–15:15 · Wrap-up

PHAROS
THE GREEK AI FACTORY

Deployment considerations

Preparing a RAG pipeline for real-world use

Deployment view of a RAG application

- A notebook proves the pipeline; deployment requires service boundaries and operational safeguards
- Key production concerns: latency, cost, data privacy, source updates, logging and regression testing

Deployment

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

Greek-language public-service RAG: practical constraints

Language
Morphology, syntax, Greek terminology and mixed formal/informal wording

Documents
Tables, calls, articles, codes, appendices and structured references

Users
Questions may be vague, incomplete or expressed in everyday language

- A specialised RAG system must bridge user language and official source language

Deployment

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

End-to-end RAG design checklist

1. Define the information need and risk level
2. Prepare clean, structured and traceable documents
3. Choose chunking and metadata deliberately
4. Combine retrieval strategies when needed
5. Construct prompts that enforce grounding and citations
6. Evaluate retrieval before evaluating answers
7. Check faithfulness, groundedness and citation quality
8. Diagnose failures and iterate systematically
9. Plan deployment: latency, privacy, monitoring and updates
10. Keep regression tests for future changes

Wrap-up

The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No. 101234269 and the Greek Ministry of Digital Governance and Artificial Intelligence.

PHAROS
THE GREEK AI FACTORY

Thank you

Questions and discussion

www.pharos-aifactory.eu

EuroHPC Joint Undertaking
Co-funded by the European Union
AI Factories PHAROS GREECE
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Ψηφιακής Διακυβέρνησης και Τεχνητής Νοημοσύνης