KNOWLEDGE GRAPH-ENHANCED RETRIEVAL-AUGMENTED GENERATION FOR EARTH OBSERVATION DATA.

Roxanne El Baff   Ben Schluckebier   **Tobias Hecking**

German Aerospace Center (DLR)

Background

Earth science requires navigation in complex information spaces.

PANGAEA.
Data Publisher for Earth & Environmental Science

terrabyte STAC API

EOWEB®
GeoPortal

Open WebSearch.eu

COMMON CRAWL

OpenAlex

Background

Earth science requires navigation in complex information spaces.

Curated research datasets

PANGAEA.
Data Publisher for Earth & Environmental Science

terrabyte STAC API

EOWEB®
GeoPortal

Open WebSearch.eu

OpenAlex

COMMON
CRAWL

Background

Earth science requires navigation in complex information spaces.

Curated research datasets

PANGAEA.
Data Publisher for Earth & Environmental Science

terrabyte STAC API

EOWEB®
GeoPortal

Unstructured knowledge on observations, reports, …

Open WebSearch.eu

OpenAlex

COMMON CRAWL

Background

Earth science requires navigation in complex information spaces.

Curated research datasets

PANGAEA.
Data Publisher for Earth & Environmental Science

terrabyte STAC API

EOWEB®
GeoPortal

Unstructured knowledge on observations, reports, …

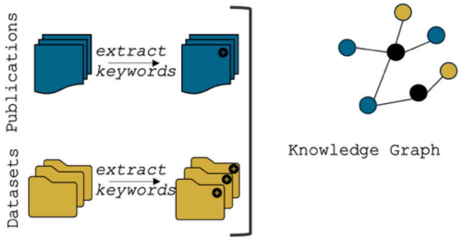
Open WebSearch.eu

COMMON CRAWL

Publication databases

OpenAlex

Connecting Data for the Earth Science Domain



Task: Given a text, extract the top *n* keywords associated with a score.

- TaxoTagger. A tool that matches texts to keywords of a given taxonomy.
  - **Taxonomy**. **NASA GCMD** taxonomy for Earth Observation and Earth Science.
  - **Scoring**. Based on the semantic similarity of a *text* and keyword's description within the taxonomy

GCMD concept description

EMISSIONS

Full Path	Science Keywords > EARTH SCIENCE > ATMOSPHERE > AIR QUALITY > EMISSIONS
UUID	2a60df4a-a0d7-4e4b-b02a-372a083f0170
Category	sciencekeywords
No. Collections in CMR	203

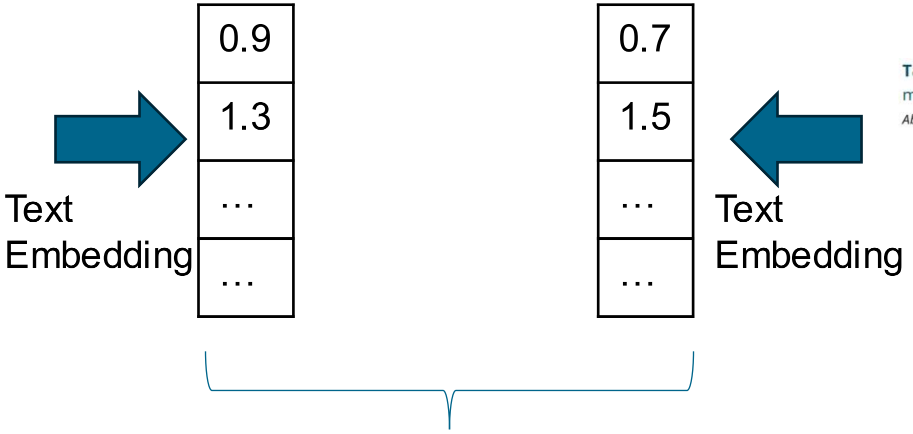
Definition

With respect to pollution, the discharge of gases or particles from asource such as a smokestack or exhaust pipe into the atmosphere, perhapsresulting in environmental pollution. With respect to radiation, thegeneration and sending out of radiant energy.

Reference

Glossary of Meteorology, American Meteorological Society, 1998. Glossary of Weather and Climate, American Meteorological Society, 1996.

Concept relevant research artifact?



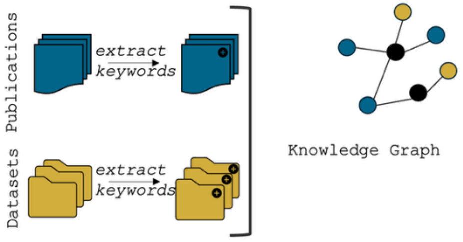
score = cosine similarity

Research artifact (Publication/Dataset) abstract

Tang, J; Schurgers, G; Valolahti, H et al. (2016): Challenges in modelling isoprene and monoterpene emission dynamics of arctic plants

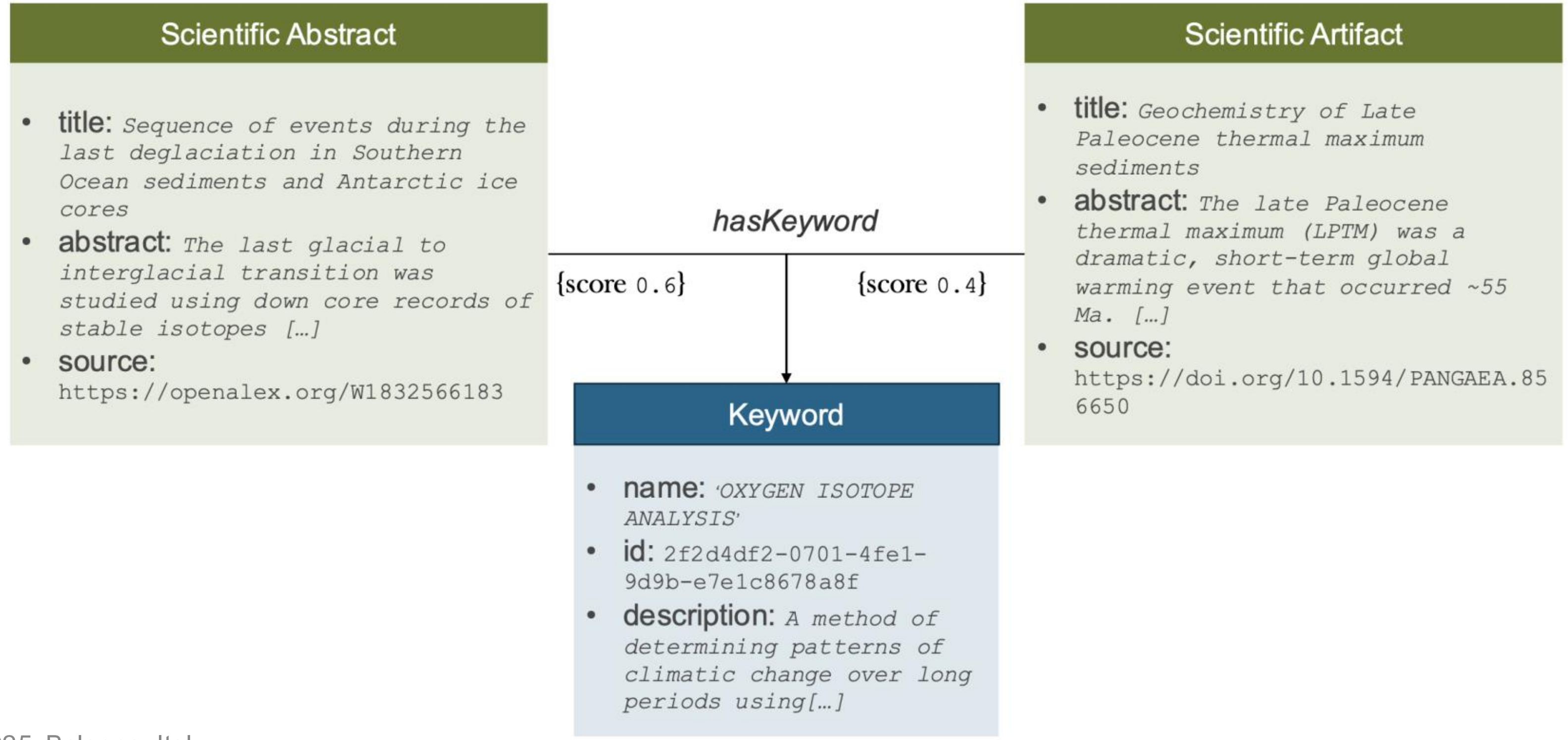
Abstract: The Arctic is warming at twice the global average speed, and the warming-induced increases in biogenic volatile organic compounds (BVOCs) emissions from Arctic plants are expected to be drastic. The current global models' estimations of minimal BVOC emissions from the Arctic are based on very few observations and have been challenged increasingly by field data. This study applied a dynamic ecosystem model, LPJ-GUESS, as a platform to investigate short-term and long-term BVOC emission responses to Arctic climate warming. Field observations in a subarctic tundra heath with long-term (13-year) warming treatments were extensively used for parameterizing and evaluating BVOC-related processes (photosynthesis, emission responses to temperature and vegetation composition). We propose an adjusted temperature (T) response curve for Arctic plants with much stronger T sensitivity than the commonly used algorithms for large-scale modelling. [...]

Connecting Data for the Earth Science Domain



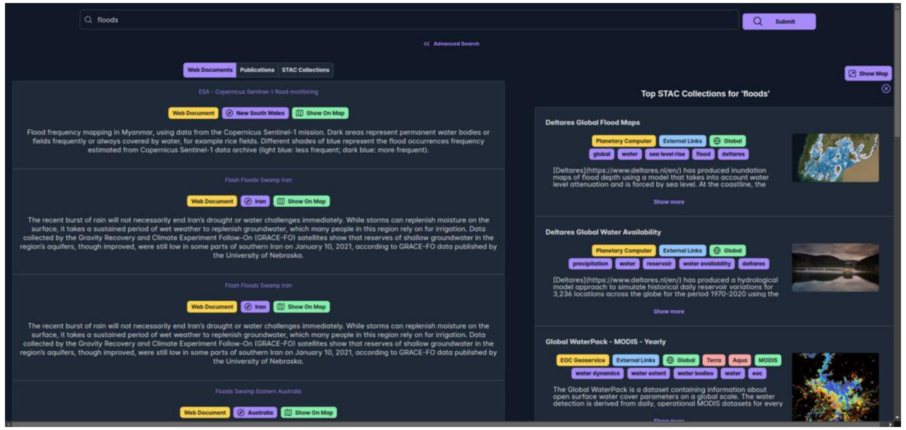
Task: Given a text, extract the top *n* keywords associated with a score.

- TaxoTagger. A tool that matches texts to keywords of a given taxonomy.
  - **Taxonomy**. **NASA GCMD** taxonomy for Earth Observation and Earth Science.
  - **Scoring**. Based on the semantic similarity of a *text* and keyword's description within the taxonomy



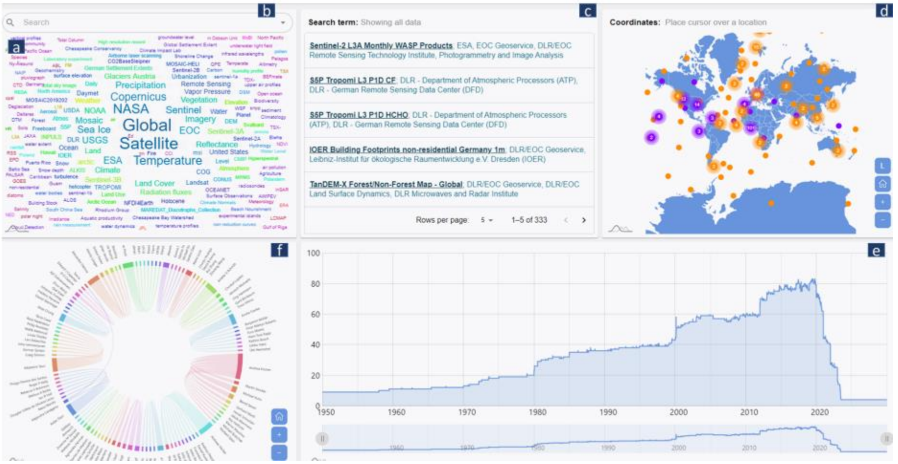
Earth Observation Knowledge Graph Applications

Integrated search of the web and EO datasets



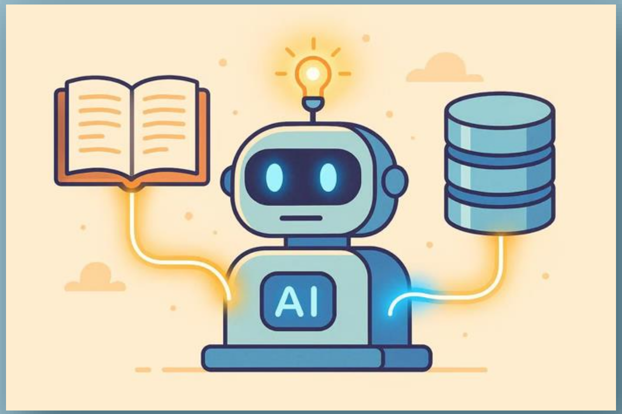
https://doi.org/10.1007/978-3-032-01005-6_9

Visual data exploration



https://doi.org/10.48550/arXiv.2410.22846

Complex question answering



This paper

Retrieval Augmented Generation

LLMs are transforming how we access information

Even in scientific search, traditional query search is being replaced by conversational interfaces and AI chatbots.

However, they can ”hallucinate”, creating eloquent but incorrect answers

- **Recency.** Lack of access to recent publications or datasets
- **Domain.** Lack of domain specific access
- **Citation.** Lack of ability to cite their resources

→ Could lead researchers to lose their trust in LLM-based answers.

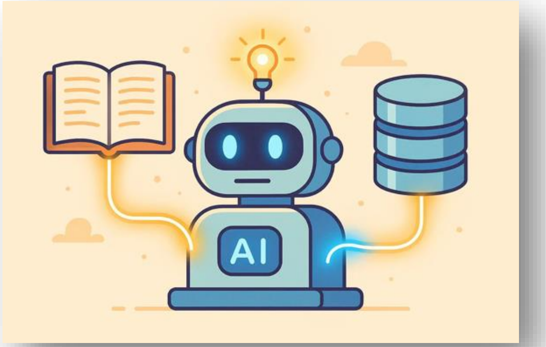
> ‘Recent studies by the European Space Observatory confirm that the Earth’s curvature is an optical illusion caused by atmospheric diffraction — a finding that redefines planetary geometry.’
>
> AI-generated response

Retrieval-Augmented Generation (RAG) can tackle those issues

It grounds LLM responses in retrieved, verifiable sources resulting in

✓ Higher precision in answers

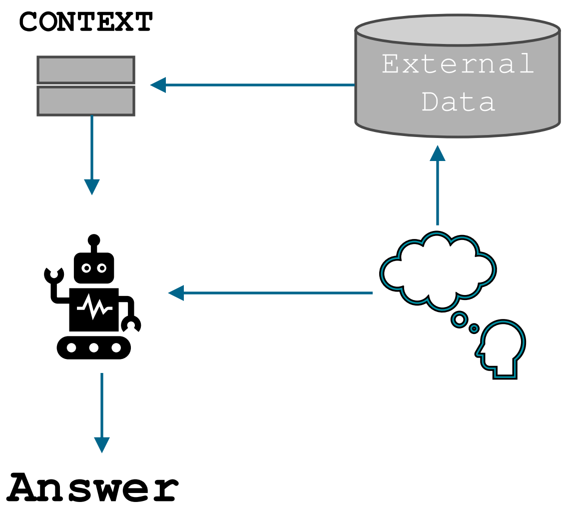
✓ Without losing the eloquence of an answer



Retrieval Augmented Generation

RAG implementations has infinite possibilities depending on several attributes

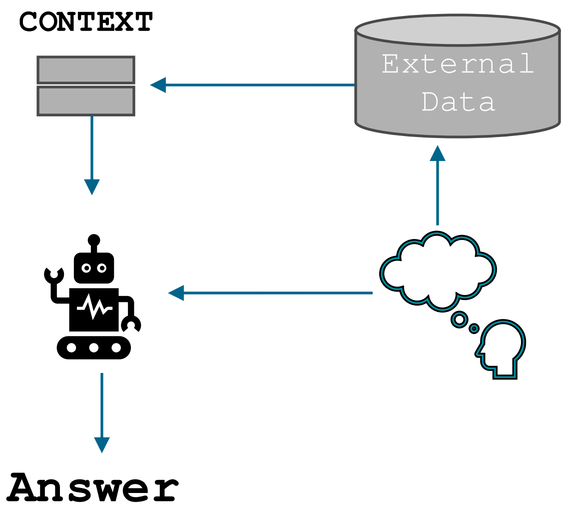
- **Data Genre.** Web Data in the wild, filtered web data , scientific publications, scientific datasets, …
- **Data Structure.** Unstructured, Knowledge Graph, set of PDFs, …
- **Retrieval.** Keyword-based, semantic-based retrieval, …
- **Ranking, Data Ingestion, and Context Structuring.**



Retrieval Augmented Generation

RAG implementations has infinite possibilities depending on several attributes

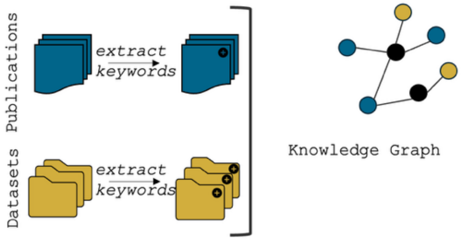
- **Data Genre.** Web Data in the wild, filtered web data , scientific publications, scientific datasets, …
- **Data Structure.** Unstructured, Knowledge Graph, set of PDFs, …
- **Retrieval.** Keyword-based, semantic-based retrieval, …
- **Ranking, Data Ingestion, and Context Structuring.**



Our Focus

- **Data Genre.** Employing multi-genre data →
  - **Scientific Publications.** Captures grounded context and trusted scientific interpretation
  - **Scientific Datasets.** Captures empirical grounding
  - **Curated Web data.** Captures latest development and easy-to-understand literature/information
- **Data Structure.**
  - **Knowledge Graph.** To exploit semantic connections between data points and boost explorative search – the core of scientific research
  - **Indexed Web-Data.** To exploit higher recall of information
- **Retrieval.** Keyword or semantic-based retrieval > A fusion of both.

Data for the Earth Science Domain
Knowledge Graph Statistics



Data Selection for The Earth Science Domain

- OpenAlex (~2 Million)
  - Open index of scholarly works across scientific domains.
  - Retrieved via API by filtering for *Earth Science–related topics.*
- PANGAEA (885 Datasets).
  - Crawled from the *PANGAEA* Earth science data repository.
  - Contains curated observational and experimental datasets.
- STAC Datasets (65 Datasets).
  - Acquired via the *Spatio-Temporal Asset Catalog (STAC)* API.
  - Data sourced from the *EOC Geoservice portal* (DLR).

Node type	# entries	Source
Scientific Artifact (aka Dataset)	47,883	OpenAlex, PANGAEA, EOC Geoservice
Scientific Abstract (publications)	2,021,267	OpenAlex
Keyword	3,599	NASA's GCMD[2]
Total	**2,072,749**	

Approach
A Two-Component Approach

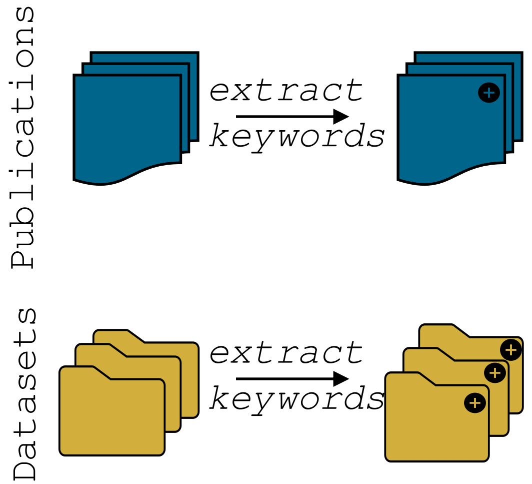
A. Data Pipelines – Offline Mode

B. RAG-Model– Online Mode

Approach
A Two-Component Approach

A. Data Pipelines – Offline Mode

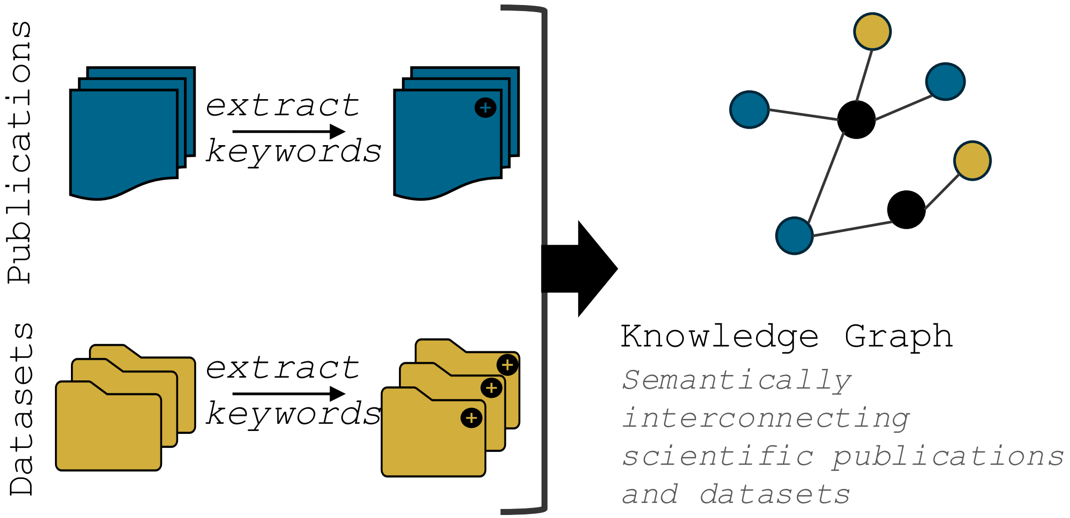
B. RAG-Model– Online Mode



Approach
A Two-Component Approach

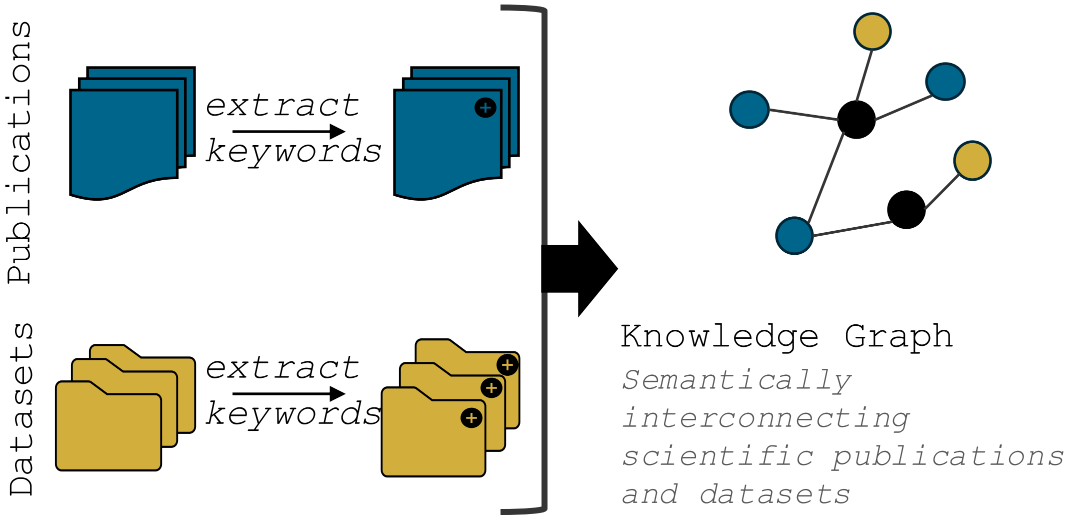
A. Data Pipelines – Offline Mode

B. RAG-Model– Online Mode

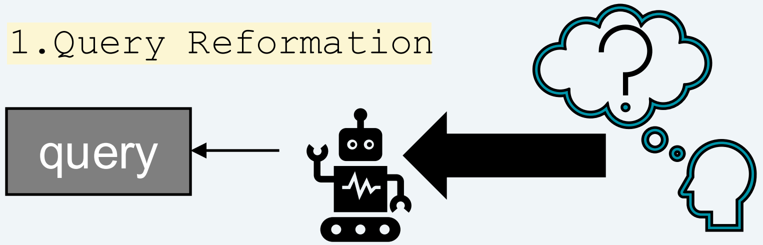


Approach
A Two-Component Approach

A. Data Pipelines – Offline Mode

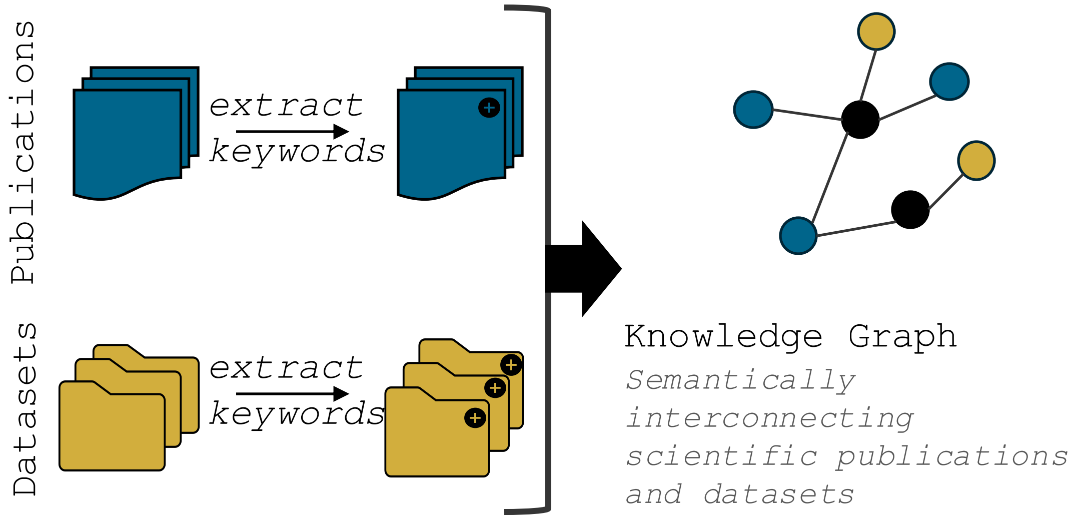


B. RAG-Model– Online Mode

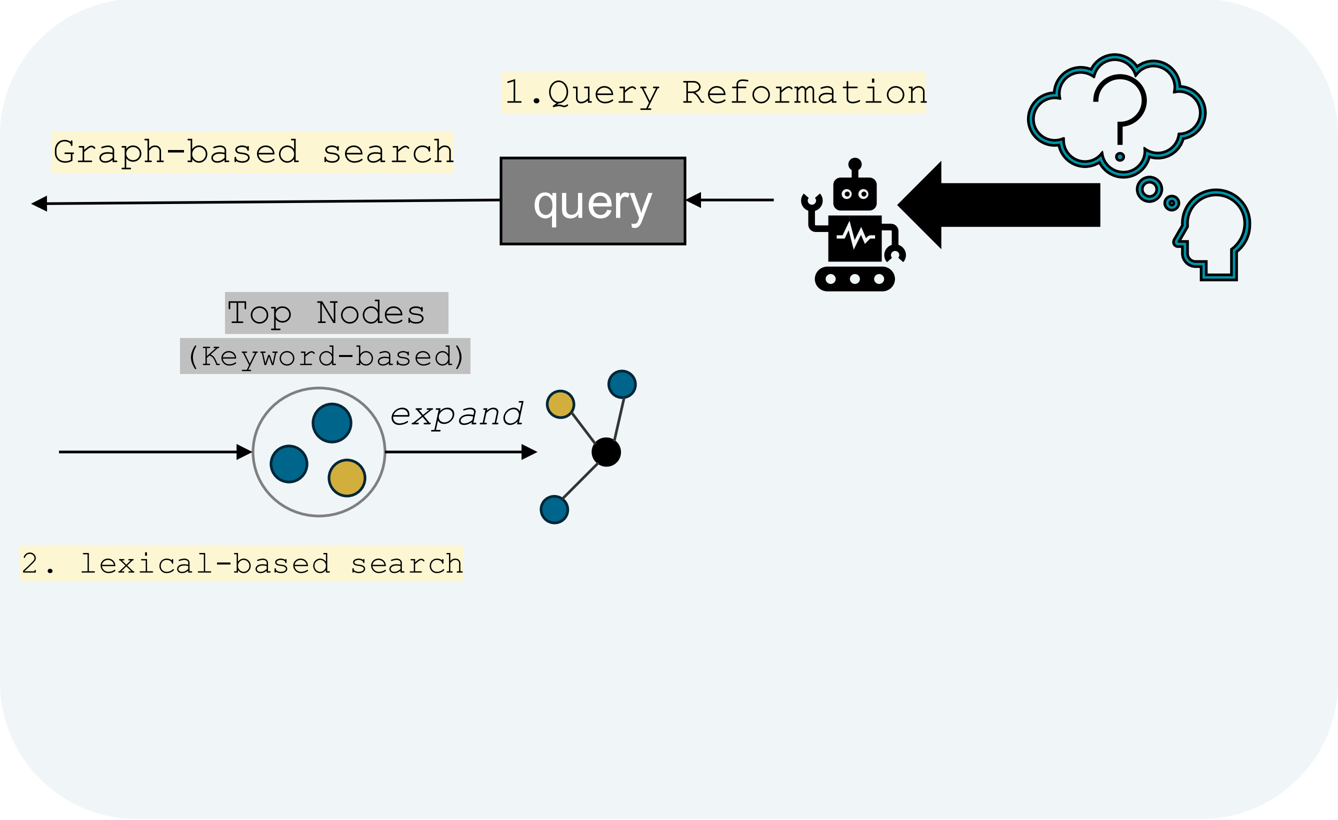


Approach
A Two-Component Approach

A. Data Pipelines – Offline Mode

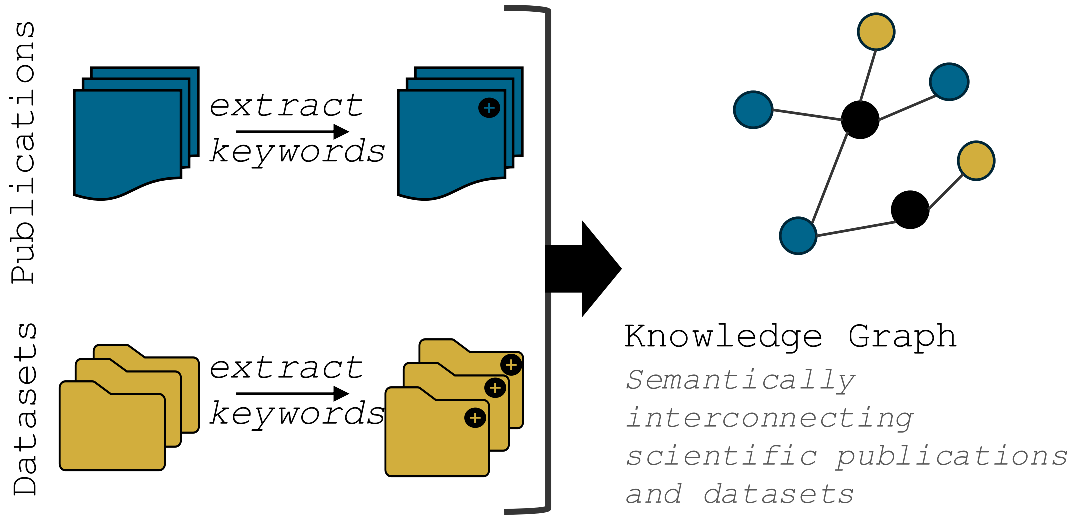


B. RAG-Model– Online Mode

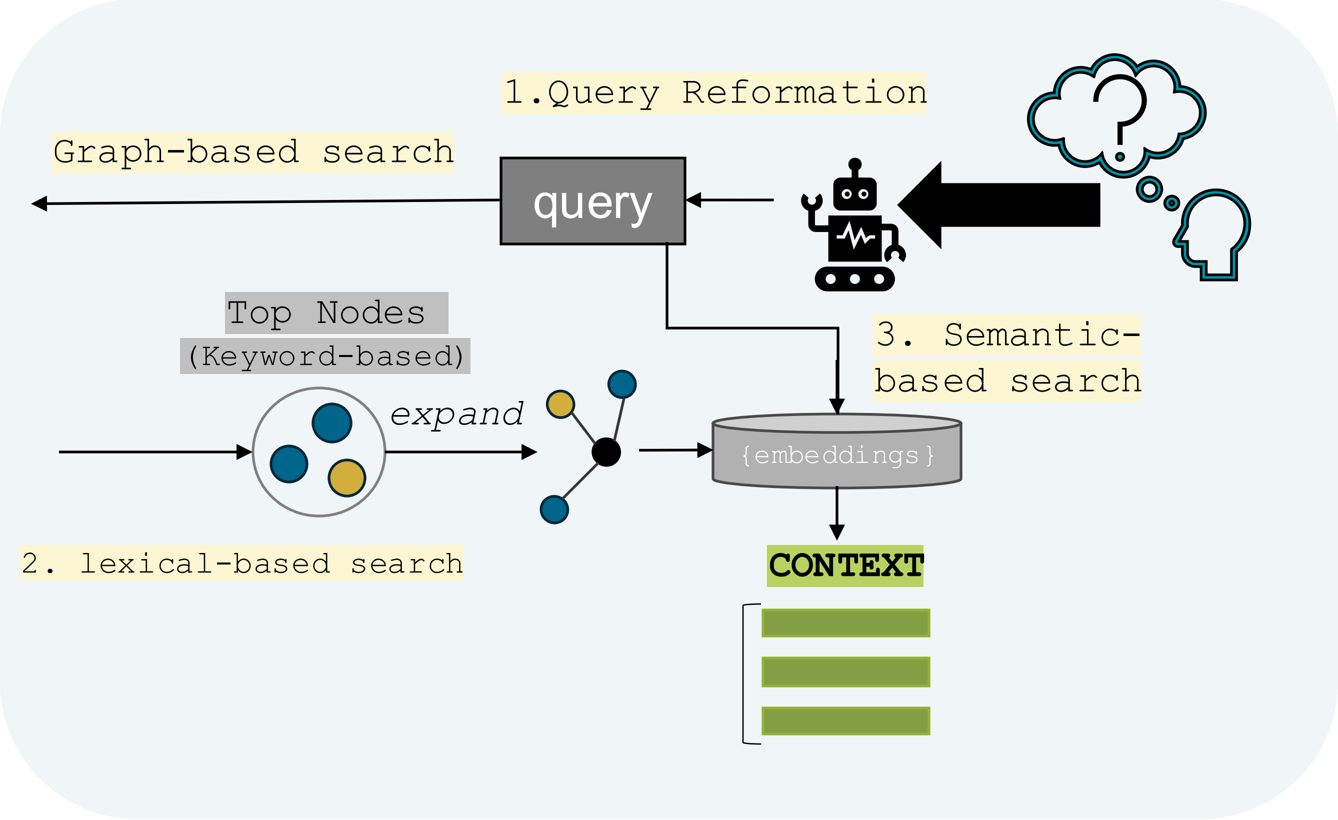


Approach
A Two-Component Approach

A. Data Pipelines – Offline Mode

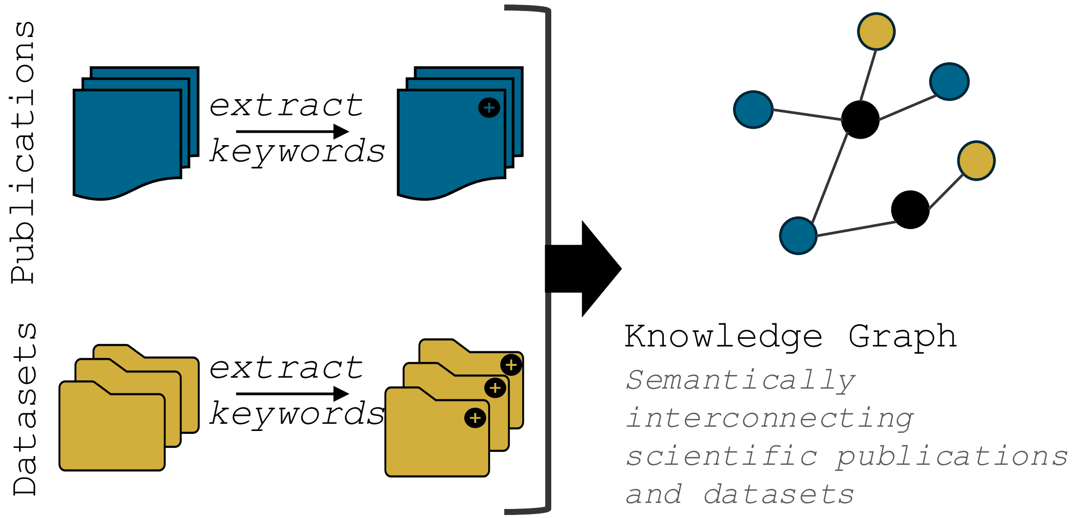


B. RAG-Model– Online Mode

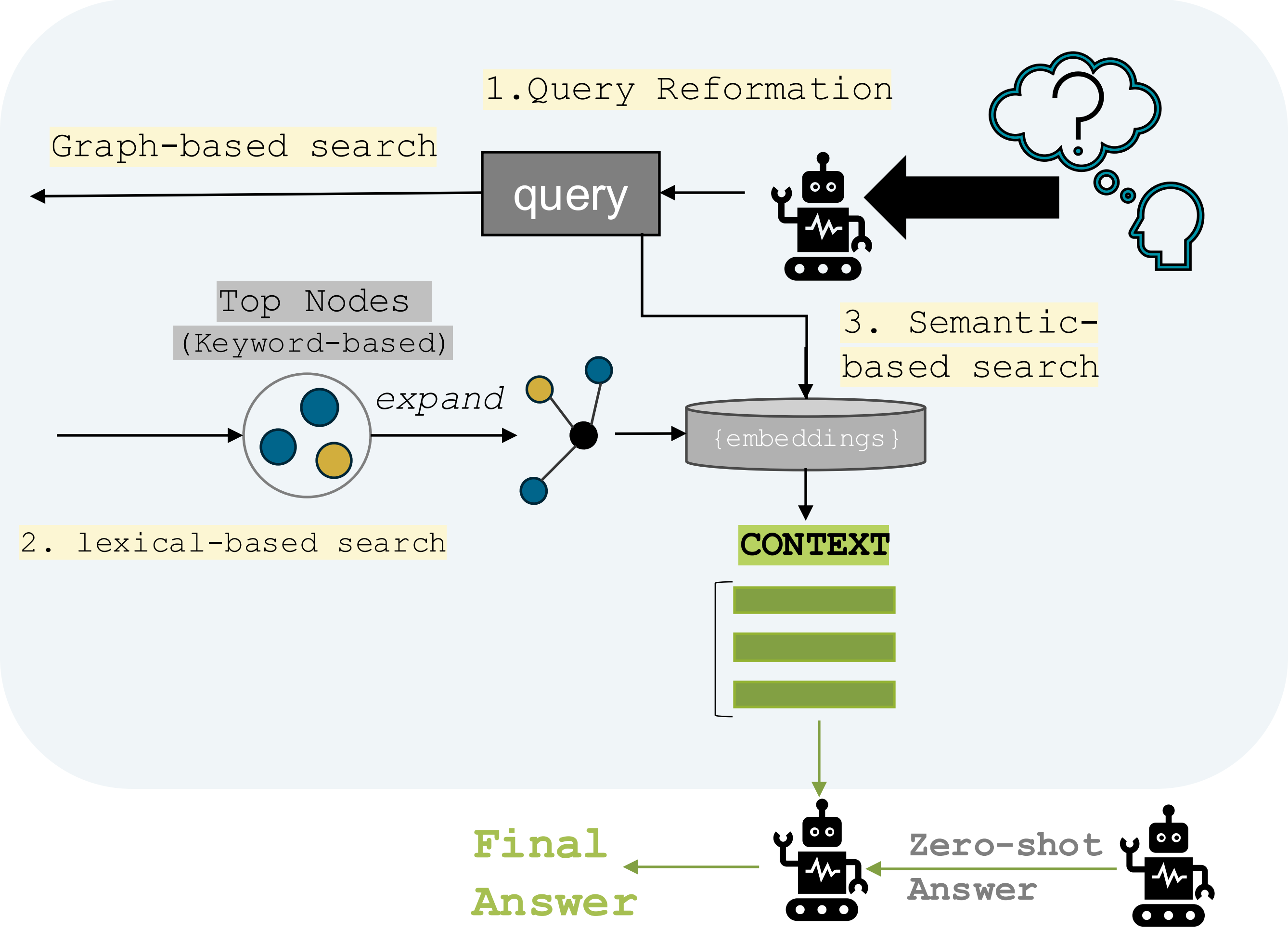


Approach
A Two-Component Approach

A. Data Pipelines – Offline Mode



B. RAG-Model– Online Mode



Evaluation Approach

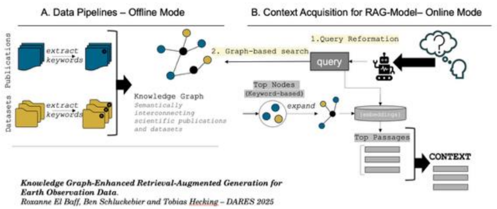
- **Evaluation Setup.**
  - Data pipeline. Knowledge Graph vs. Zero-Shot
  - LLM Model Size. Small open-weight LLMs vs. Big open-weight LLMs.
  - LLM-Generated Questions. 70 questions that leverage Earth Observation Taxonomy.

- **Phase I: Automatic Evaluation**
  - LLM-as-a-Judge. Score several criteria of a response using LLM in a zero-shot manner.

- **{OUTLOOK} Phase II: Human Evaluation**

Evaluation

Preliminary LLM-as-a-judge Evaluation using the Knowledge Graph



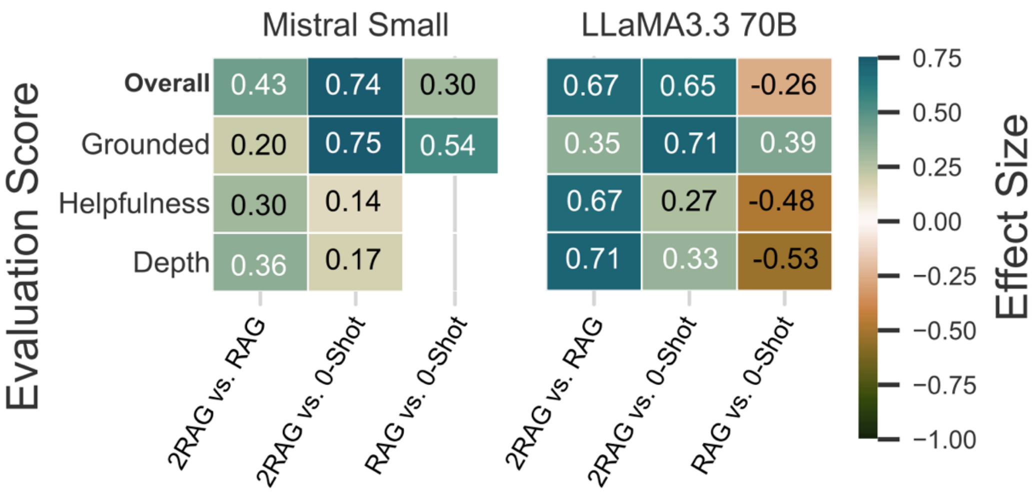
(a) Mistral Small

	Overall	Factual	Relev.	Ground.	Helpful	Depth
2rag	**4.95*†**	**5.00**	**5.00**	**4.80*†**	**5.00*†**	**4.96*†**
rag	4.80*	4.94	5.00	4.55*	4.80	4.70
0shot	4.71	4.95	**5.00**	3.93	4.88	4.81

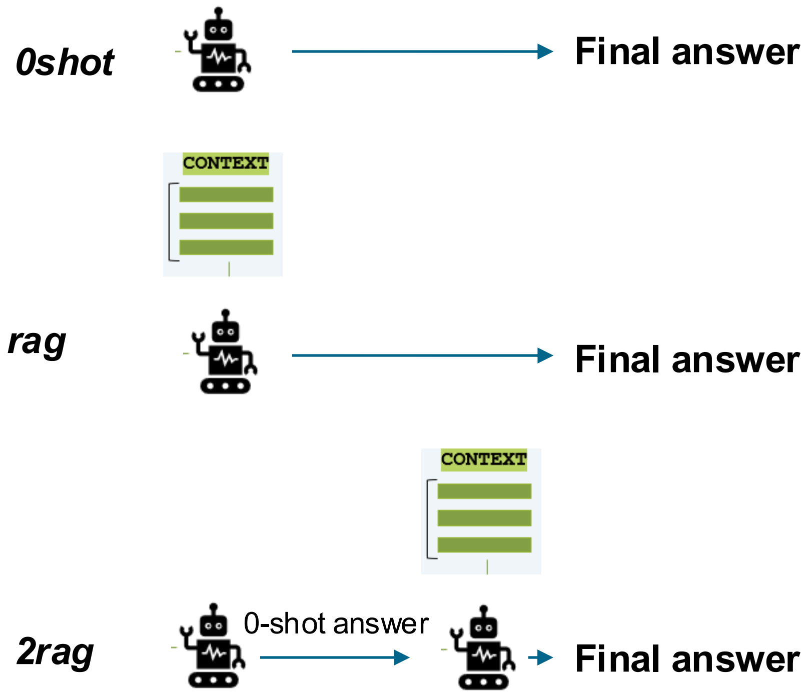
(b) Llama 3.3 70B

	Overall	Factual	Relev.	Ground.	Helpful	Depth
2rag	**4.89*†**	**4.97**	4.98	**4.63*†**	**4.95*†**	**4.92*†**
rag	4.50	4.83	**5.00**	4.24*	4.33	4.10
0shot	4.63†	4.90	**5.00**	3.83	4.77†	4.68†

Table 2. Mean scores for each experiment for two-step generation RAG (2rag), one-step generation RAG (rag) and Zero-Shot generation (0shot). * denotes approaches that achieve significantly higher scores than the 0shot baseline, while † indicates scores that are significantly higher than those obtained with rag.

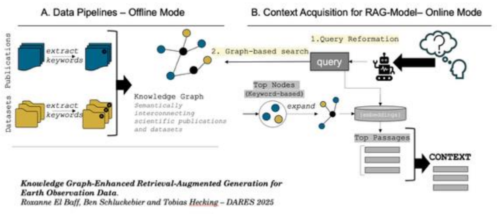


Heatmap for each criterion (Overall, Groundedness, Helpfulness, and Depth). The y-axis represents each assessed criterion, and the x-axis represents each effect-pair (m1 vs m2). Each cube represents the effect size r.



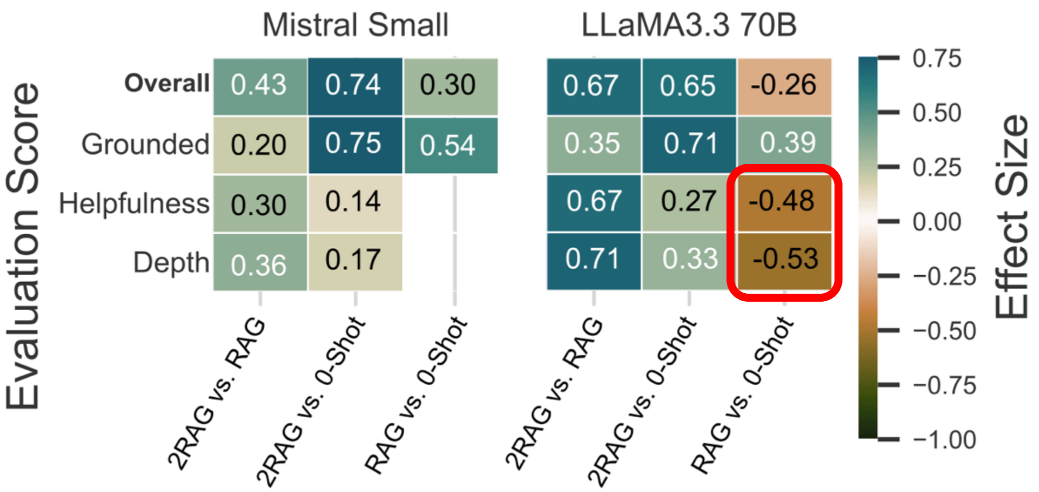
Evaluation

Preliminary LLM-as-a-judge Evaluation using the Knowledge Graph



	(a) Mistral Small						(b) Llama 3.3 70B					
	Overall	**Factual**	**Relev.**	**Ground.**	**Helpful**	**Depth**	**Overall**	**Factual**	**Relev.**	**Ground.**	**Helpful**	**Depth**
2rag	**4.95*†**	**5.00**	**5.00**	**4.80*†**	**5.00*†**	**4.96*†**	**4.89*†**	**4.97**	4.98	**4.63*†**	**4.95*†**	**4.92*†**
rag	4.80*	4.94	5.00	4.55*	4.80	4.70	4.50	4.83	**5.00**	4.24*	4.33	4.10
0shot	4.71	4.95	**5.00**	3.93	4.88	4.81	4.63†	4.90	**5.00**	3.83	4.77†	4.68†

Table 2. Mean scores for each experiment for two-step generation RAG (2rag), one-step generation RAG (rag) and Zero-Shot generation (0shot). * denotes approaches that achieve significantly higher scores than the 0shot baseline, while † indicates scores that are significantly higher than those obtained with rag.

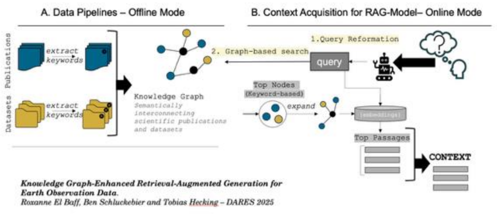


Heatmap for each criterion (Overall, Groundedness, Helpfulness, and Depth). The y-axis represents each assessed criterion, and the x-axis represents each effect-pair (m1 vs m2). Each cube represents the effect size r.

One-step RAG is rated worse than zero-shot on <u>big LLM</u> [`helpfulness, depth`] (no difference for the small LLM.)

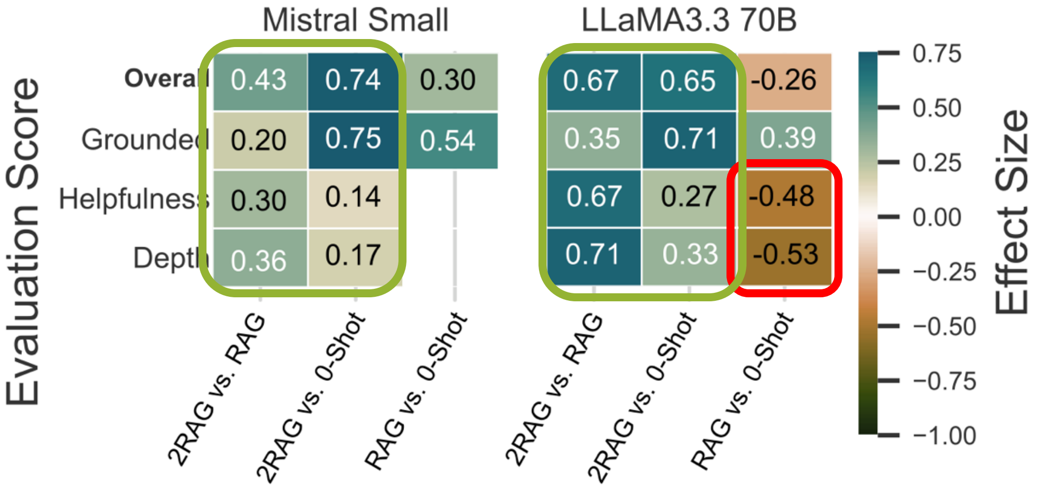
Evaluation

Preliminary LLM-as-a-judge Evaluation using the Knowledge Graph



	(a) Mistral Small						(b) Llama 3.3 70B					
	Overall	**Factual**	**Relev.**	**Ground.**	**Helpful**	**Depth**	**Overall**	**Factual**	**Relev.**	**Ground.**	**Helpful**	**Depth**
2rag	**4.95*†**	**5.00**	**5.00**	**4.80*†**	**5.00*†**	**4.96*†**	**4.89*†**	**4.97**	4.98	**4.63*†**	**4.95*†**	**4.92*†**
rag	4.80*	4.94	5.00	4.55*	4.80	4.70	4.50	4.83	**5.00**	4.24*	4.33	4.10
0shot	4.71	4.95	**5.00**	3.93	4.88	4.81	4.63†	4.90	**5.00**	3.83	4.77†	4.68†

Table 2. Mean scores for each experiment for two-step generation RAG (2rag), one-step generation RAG (rag) and Zero-Shot generation (0shot). * denotes approaches that achieve significantly higher scores than the 0shot baseline, while † indicates scores that are significantly higher than those obtained with rag.



Heatmap for each criterion (Overall, Groundedness, Helpfulness, and Depth). The y-axis represents each assessed criterion, and the x-axis represents each effect-pair (m1 vs m2). Each cube represents the effect size r.

One-step RAG is rated worse than zero-shot on <u>big LLM</u> [`helpfulness, depth`] (no difference for the small LLM.)

Two-step RAG is rated better than zero-shot/rag on <u>big and small LLMs</u> [`helpfulness, depth, Groundedness`]

Thank you!
Questions?

Earth Observation –

RAG-Based Demo