NLP-BASED COGNITIVE SEARCH ENGINE FOR THE GEOSS PLATFORM DATA

Exploiting existing GEOSS datasets for Climate Change adaption and mitigation

Scan for abstract

Sharing is encouraged

LIBRA A.I. Technologies

Eiffel GEOSS APPLICATIONS FOR CLIMATE CHANGE

Yannis Kopsinis*, Zisis Flokas*,
Pantelis Mitropoulos*, Christos Petrou*, Thodoris Siozos*, Giorgos Siokas*
**LIBRA AI Technologies, yannis.kopsinis@libramli.ai*

01. Introduction

This work presents a domain-aware cognitive search engine (SE) designed in EIFFEL H2020 project. It aims to exploit recent advances in Machine Learning (ML)-based Natural Language Processing (NLP) to overcome current challenges in the searching capabilities of Data Portals. The system includes an optimized AI Large Language Model (LLM) retrained with an extensive Climate Change (CC)-specific text corpus. Cognitive search adds language understanding to the search results, promoting the most semantically relevant results to the top. The use-case is the GEOSS Portal, but the same principles apply elsewhere.

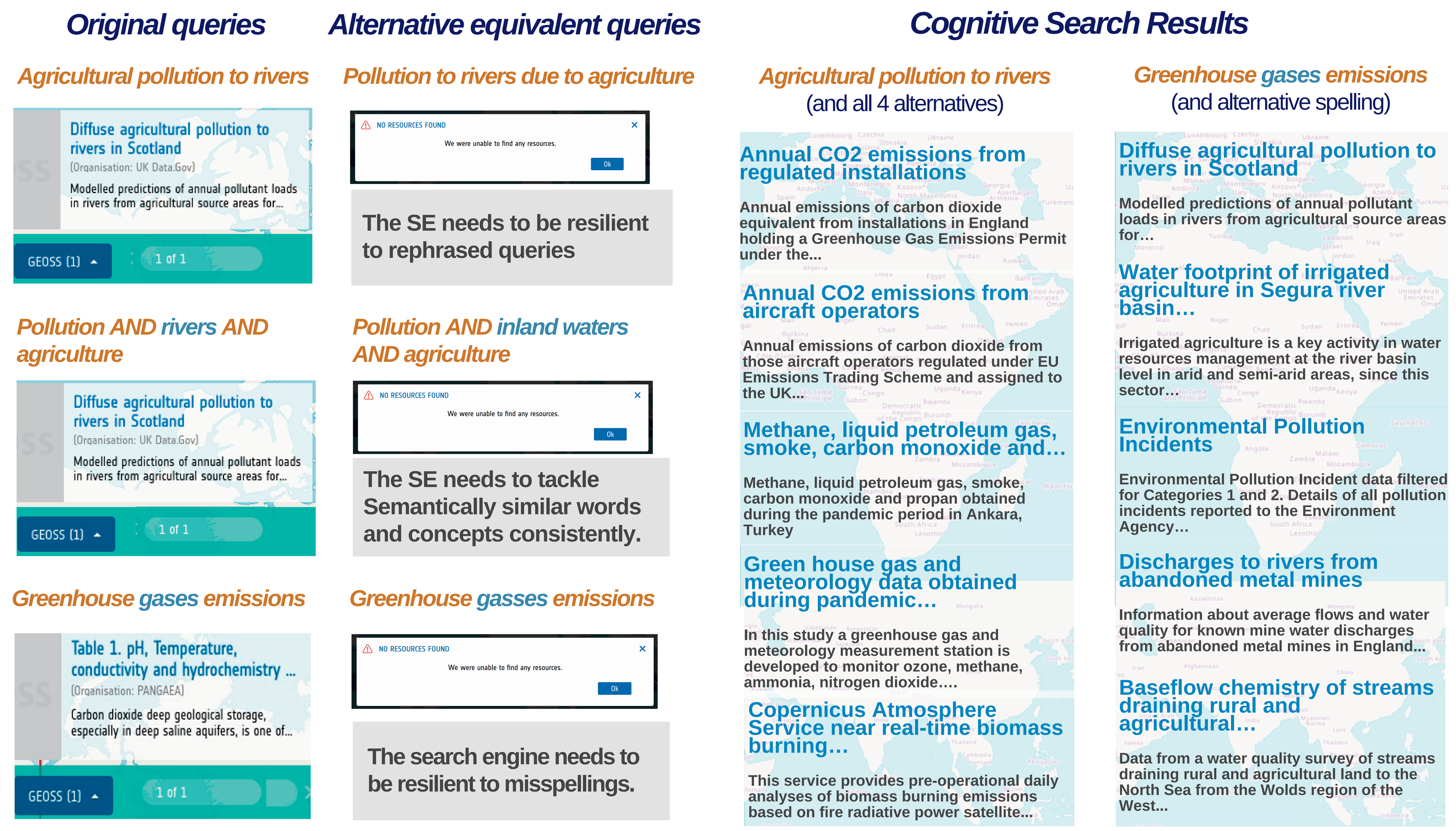
02. Conventional vs Cognitive search data discovery experience

Conventional search engines, facts

- Based on exact or fussy term searches.
- Limited data discovery capabilities.
- The inherent limitations of such approaches are expressed in a higher degree in metadata querying since the available text is limited to, usually, title, description and keyword lists.

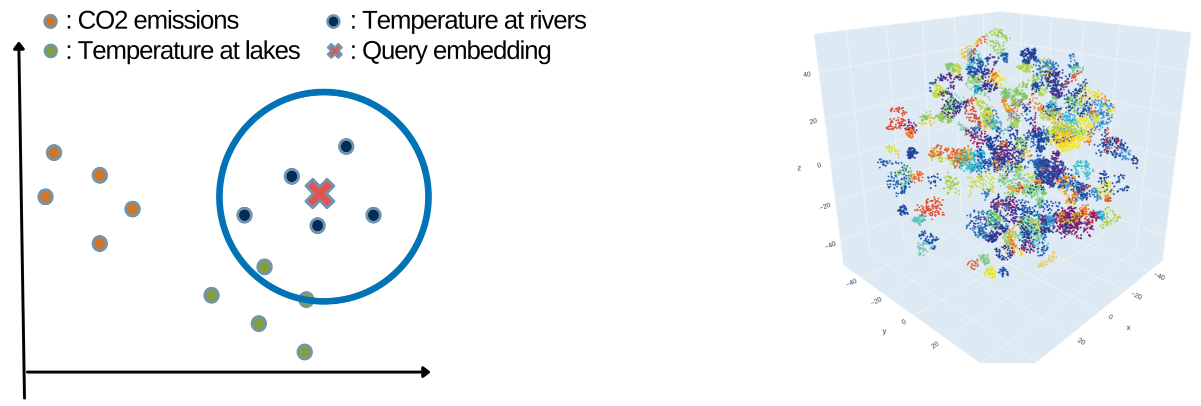
Cognitive search engines, facts

- Based on data-driven language models.
- Allow free-text querying.
- The language model inherently performs the semantic analysis: It understands individual words' meanings, the meaning of words within their context, the semantic relationships between individual words and even the meaning of whole sentences.



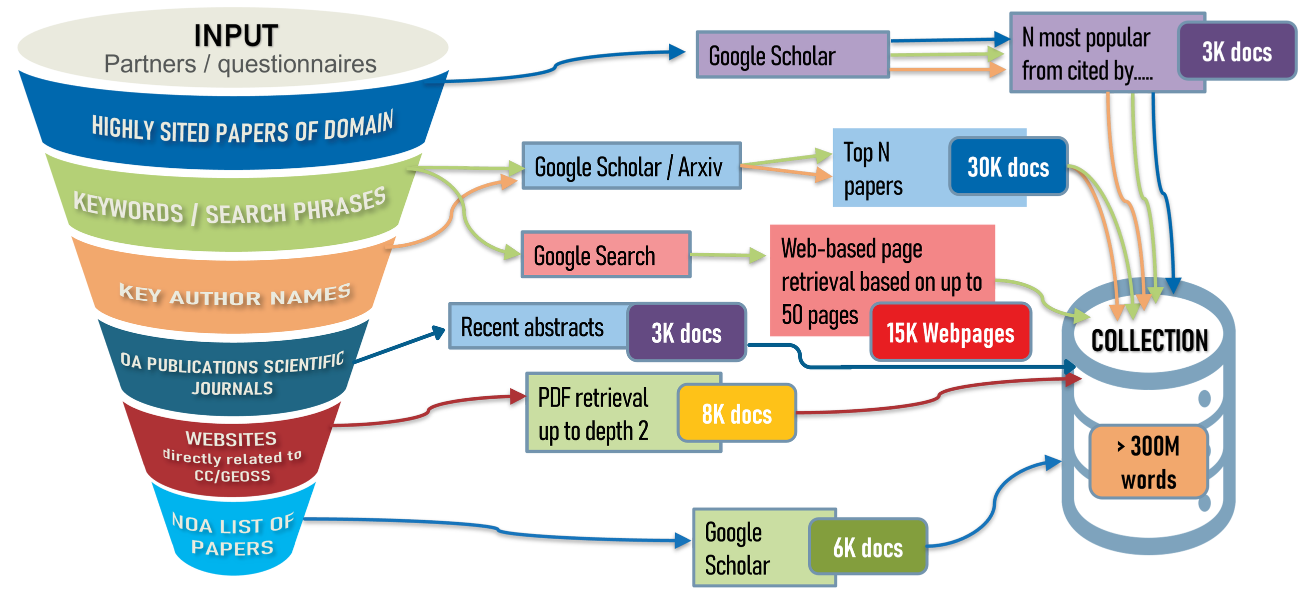
03. How does cognitive search works

- The LLM transforms all documents into mathematical vectors (**Embeddings**), inherently performing semantic analysis.
- Words/terms and sentences with similar concepts and meanings lie close in the embedding space.
- The distance between the document and the query embeddings measures the relevance between a document in the database and any query.
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top.
- It can be domain-aware: In EIFFEL we aim for CC domain specificity.



Text embedding encodes words and sentences as numeric vectors

04. CC Domain-specific corpus collection for LLM training

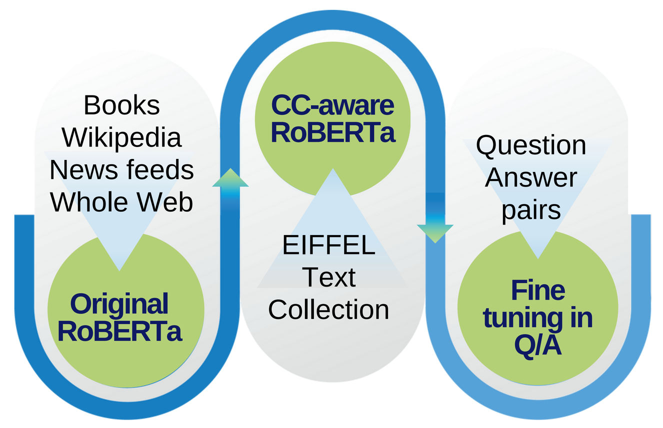


"basin", "distribution", "parameters", "factors", "regions", "environmental", "variables", "emissions", "simulation", "atmospheric", "correlation", "modelling, "measurement", "estimation", "greenhouse", "radiation", "percentage", "climatic", "cooling", "rainfall", "regression", "gases", "pollution", "meteorological", "dioxide", "flux", "anthropogenic", "indicator", "humidity", "ocean", "baseline", "ecosystems", "renewable", "hydrological", "sustainable", "socioeconomic", "CO_2"

CC-related document collection process for LLM retraining (13M sentence) and newly included terms in the Large Language Model

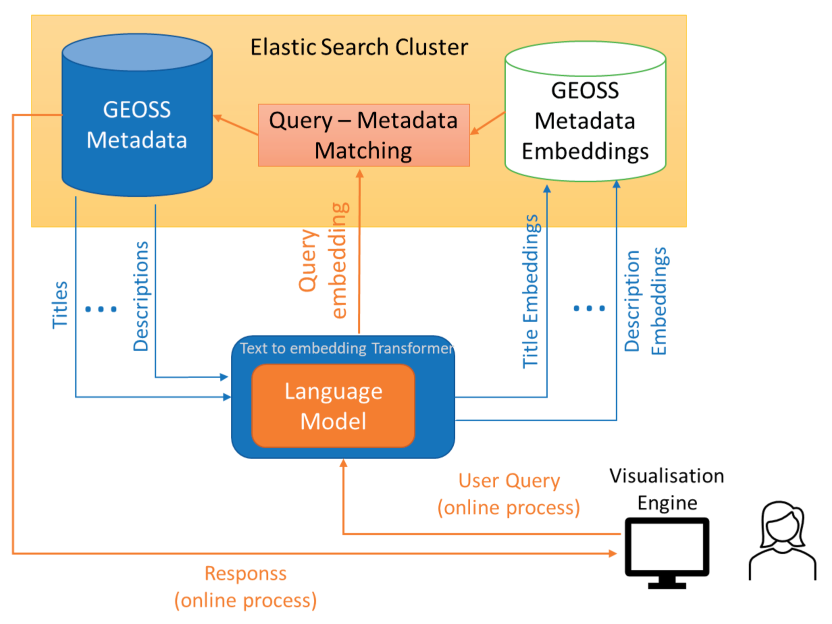
05. LLM fine-tuning for domain-aware cognitive search

- **Step A**: Unsupervised learning with the CC-related corpus that includes new terms[1].
- **Step B**:
  - We use an independent, instruction-based LLM (such as chatGPT) to generate Q&A pairs from the CC-related corpus and the GEOSS Portal metadata description field.
  - We use this new dataset for finetuning in the domain using Generative Pseudo Labeling (GPL)[2] approach.
- **Alternative path:** The Q&A pairs dataset is used for supervised training of a dedicated Cross-encoder (work in progress).



06. EIFFEL Cognitive search pipeline

- The metadata (e.g., titles, descriptions, keywords) pass through the LLM to produce metadata embeddings (offline process).
- The user query passes through the language model to produce the query embedding (online process).
- The semantically similar data objects are returned in ranked order.
- Elasticsearch stores embeddings and calculates vector similarity fast.



References

1.*Wang, Kexin, Nils Reimers, and Iryna Gurevych. "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoderfor Unsupervised Sentence Embedding Learning." Findings of the Association for Computational Linguistics: EMNLP 2021.*
2. *Wang, Kexin, et al. "GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Acknowledgment

This work is part of the EIFFEL EU project. The EIFFEL project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101003518.