Large Language Models and SDMX: From Natural Language to Structured Stats Navigation

Alessandro Benedetti, Director @ Sease

2024 SDMX Experts Workshop - *07/10/2024*

Sease | Information Retrieval Applied

WHO AM I ?

ALESSANDRO BENEDETTI

- Born in **Tarquinia** (ancient Etruscan city in Italy)
- **R&D Software Engineer**
- Director
- Master degree in **Computer Science**
- PC member for **ECIR**, **SIGIR** and **Desires**
- **Apache Lucene/Solr PMC member/committer**
- Elasticsearch/OpenSearch expert
- Semantic search, NLP, Machine Learning technologies passionate
- Beach Volleyball player and Snowboarder

SEArch SEervices

- Headquarter in **London**/distributed
- **Open-source** Enthusiasts
- **Apache Lucene/Solr** experts
- **Elasticsearch/OpenSearch** experts
- Community **Contributors**
- Active **Researchers**

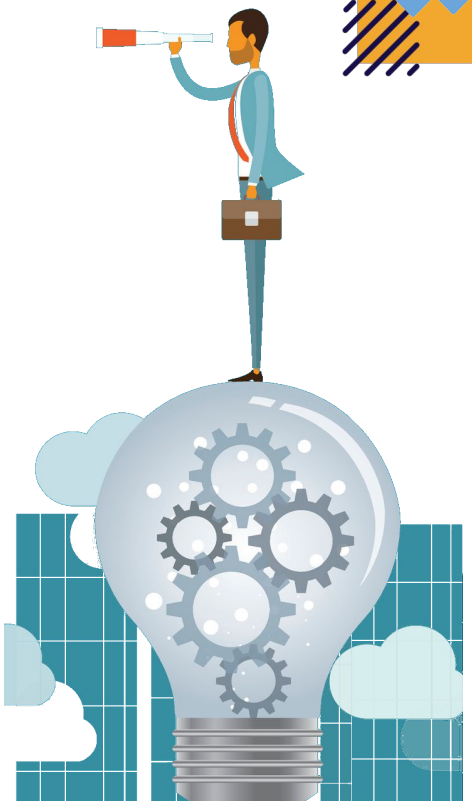
HOT TRENDS:

- **Large Language Models** Applications
- **Vector-based (Neural) Search**
- Natural Language Processing
- **Learning To Rank**
- Document Similarity
- **Search Quality Evaluation**
- Relevance Tuning

www.sease.io

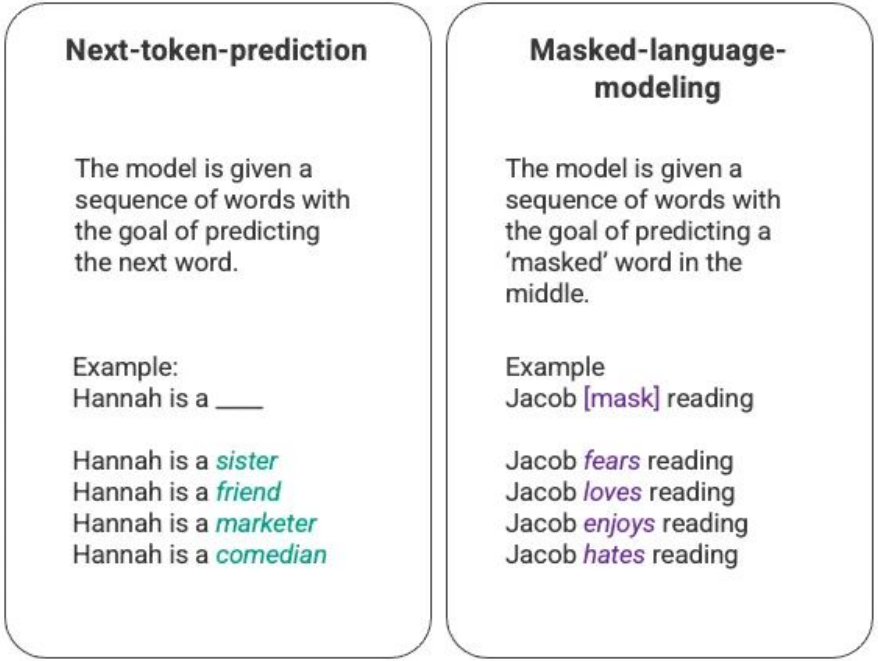
AGENDA

- Use Case Overview
- From Natural Language to Structured Queries
- Findings
- The Road to Production



WHAT IS A LARGE LANGUAGE MODEL

- **Next-token-prediction** and **masked-language-modeling**
- Estimate the likelihood of each possible word (in its vocabulary) given the previous sequence
- Learn the statistical structure of language
- Pre-trained on huge quantities of text
- Fine-tuned for different tasks (**Following Instructions**)



LEXICAL PROBLEMS

VOCABULARY MISMATCH PROBLEM

- **Terms matching** between the query and the documents.
  - **false positive**: docs retrieved (terms match) but the information need is not satisfied
  - **false negative**: docs not retrieved (terms don't match) but there was the information need in the corpus → **zero result query**

SEMANTIC SIMILARITY

- **Same terms different meaning**: How old are you? - How are you?
- **Different terms same meaning**: How old are you? - What is your age?

DISAMBIGUATION

- Same term in two totally different contexts assume totally different meanings

LEXICAL SOLUTIONS

There are some **lexical solutions** to these:

Manually curated

- Synonyms, Hypernyms, Hyponyms

Algorithmic

- Stemming, lemmatization
- Knowledge Base disambiguation

These solutions are expensive to maintain and do not guarantee high quality results.
We can do better!

NATURAL LANGUAGE QUERY PARSING

PM10 levels produced by industries in the European Community in May 2015

```
{
  "filters": {
      "Country": "European Union (28 countries)#EU28#”,
      "Pollutant": "Particulates (PM10)#PM10#",
      "Variable": "Total man-made emissions#TOT#|Industrial combustion#STAT_COMB_IND#",
      "Time Period": "Second trimester(Q2)",
      "Year": "2015"
  }
}
```

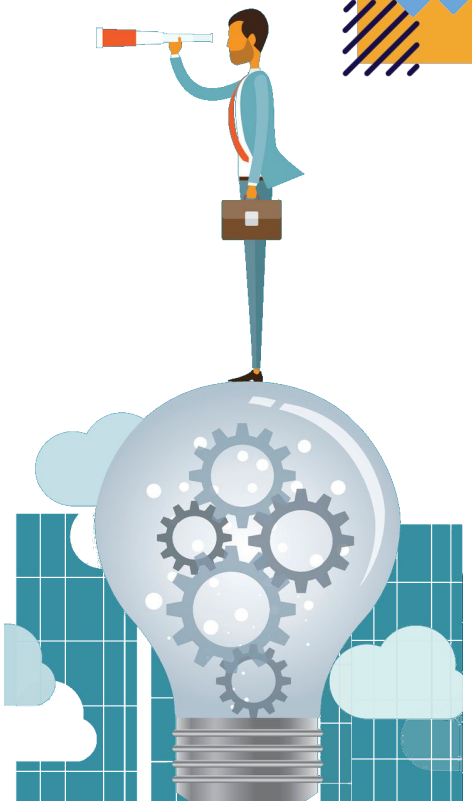
REAL CASE APPLICATION

We have been working with SDMX sponsor organisations to exploit a LLM in order to:

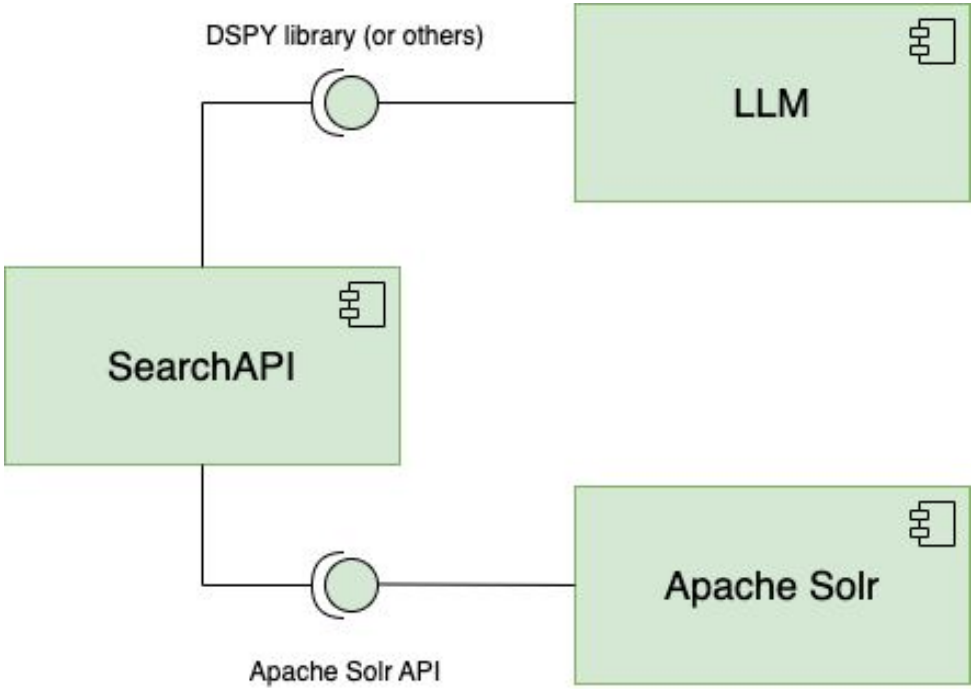
- **Disambiguate** the meaning of a user's natural language query
- **Extract** the relevant information from it
- Use the extracted information to **implement a structured Solr query**

AGENDA

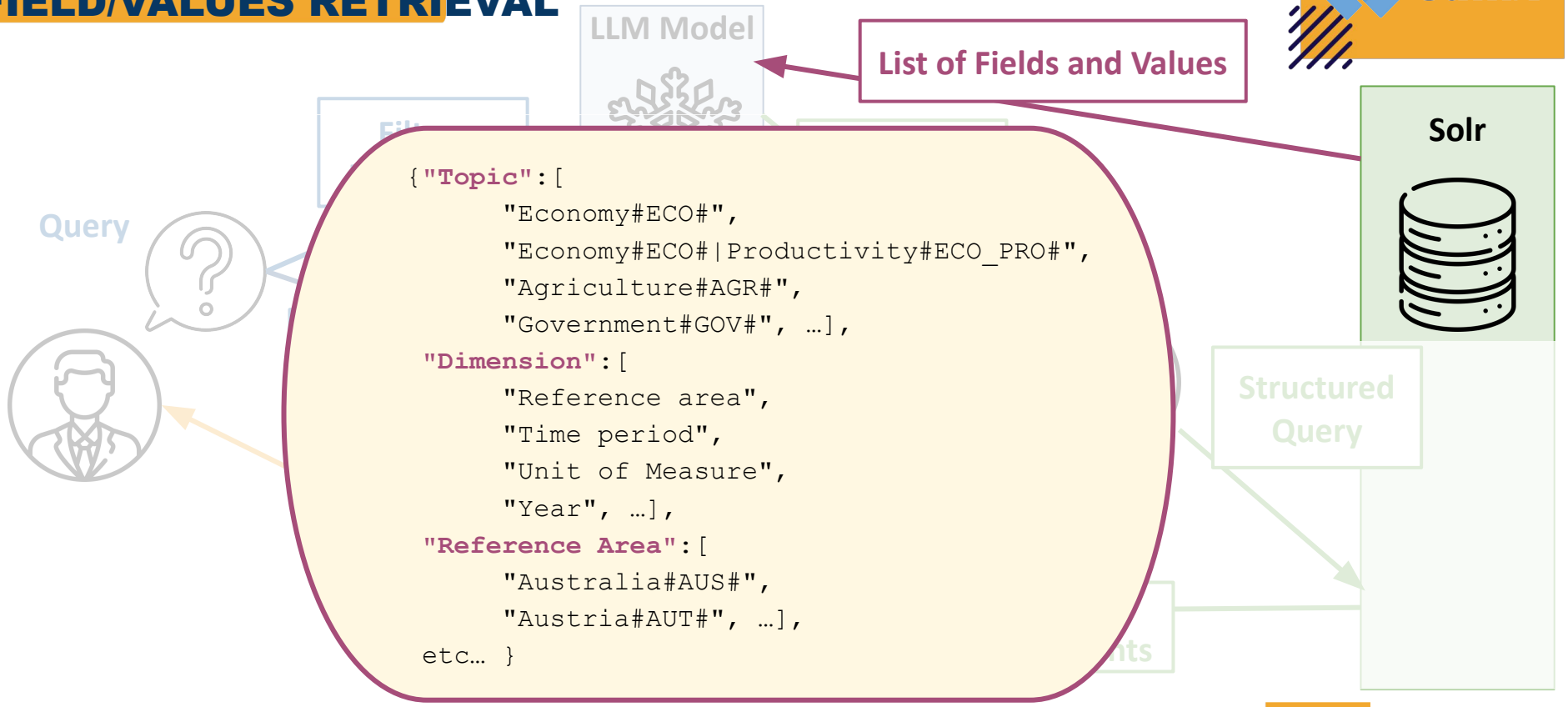
- Use Case Overview
- From Natural Language to Structured Queries
- Findings
- The Road to Production



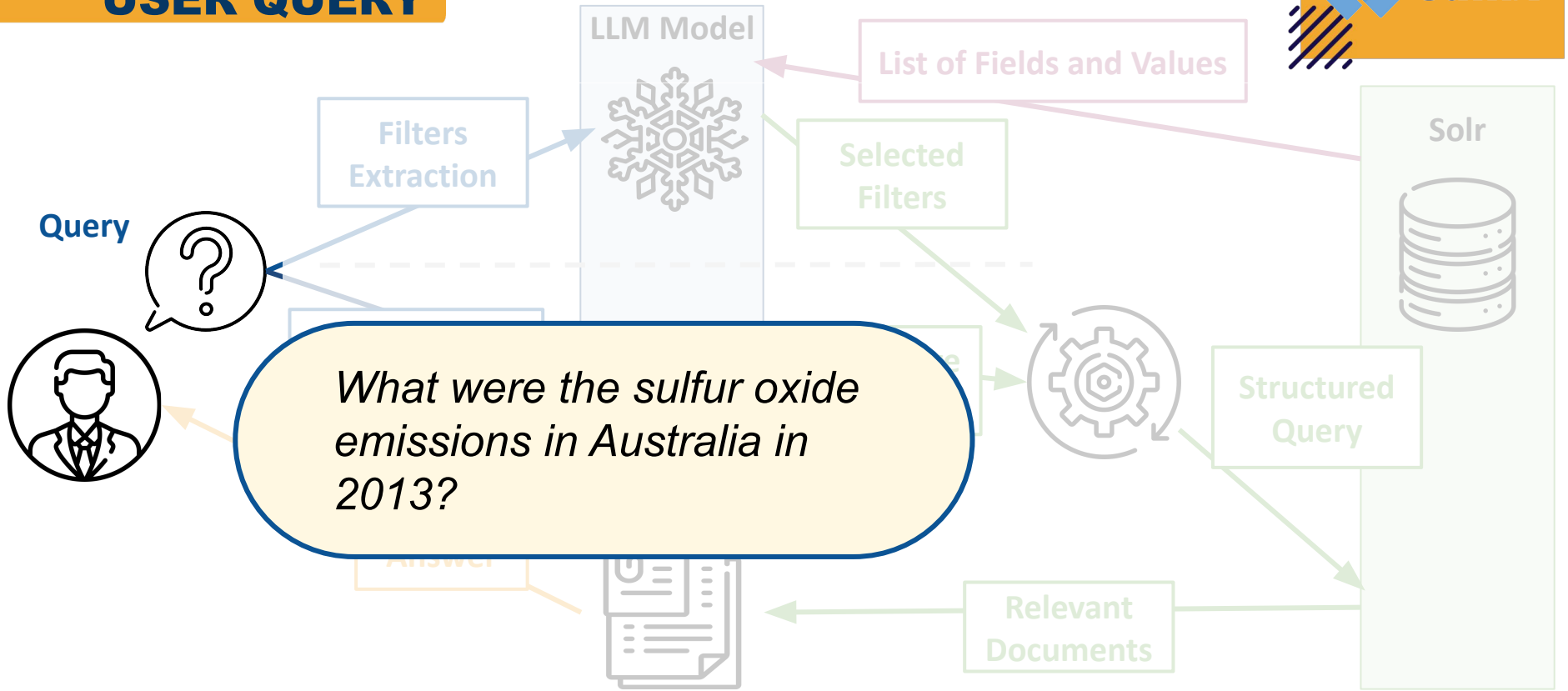
ARCHITECTURE



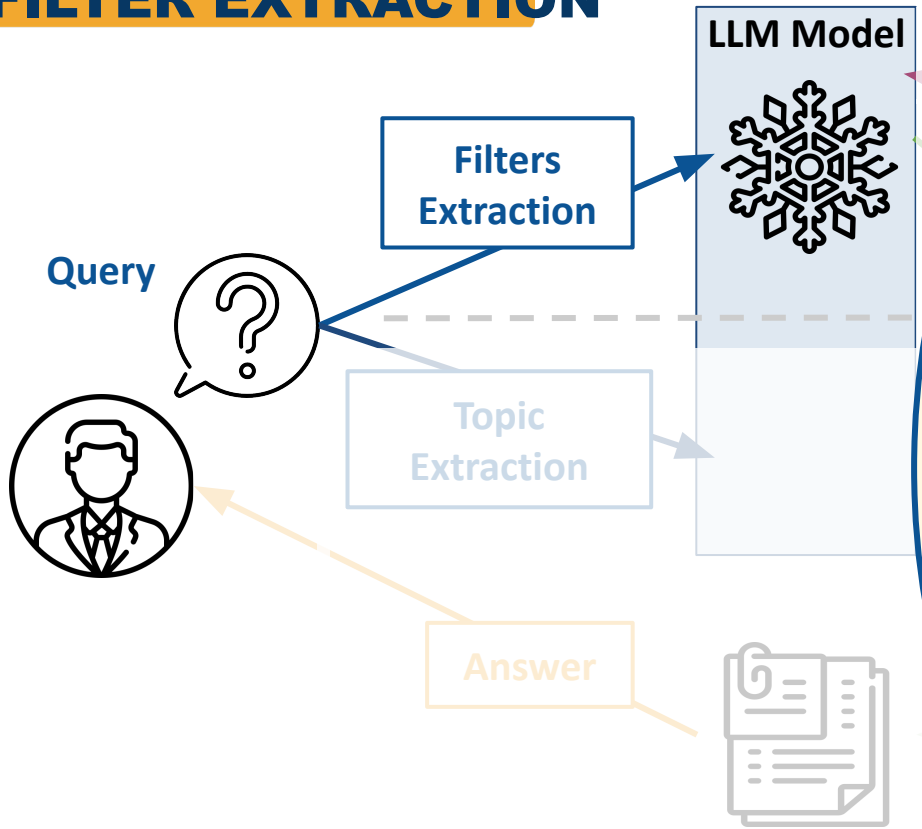
FIELD/VALUES RETRIEVAL



USER QUERY


LLM Model
List of Fields and Values
Filters Extraction
Selected Filters
Solr
Query
What were the sulfur oxide emissions in Australia in 2013?
Structured Query
Relevant Documents

FILTER EXTRACTION



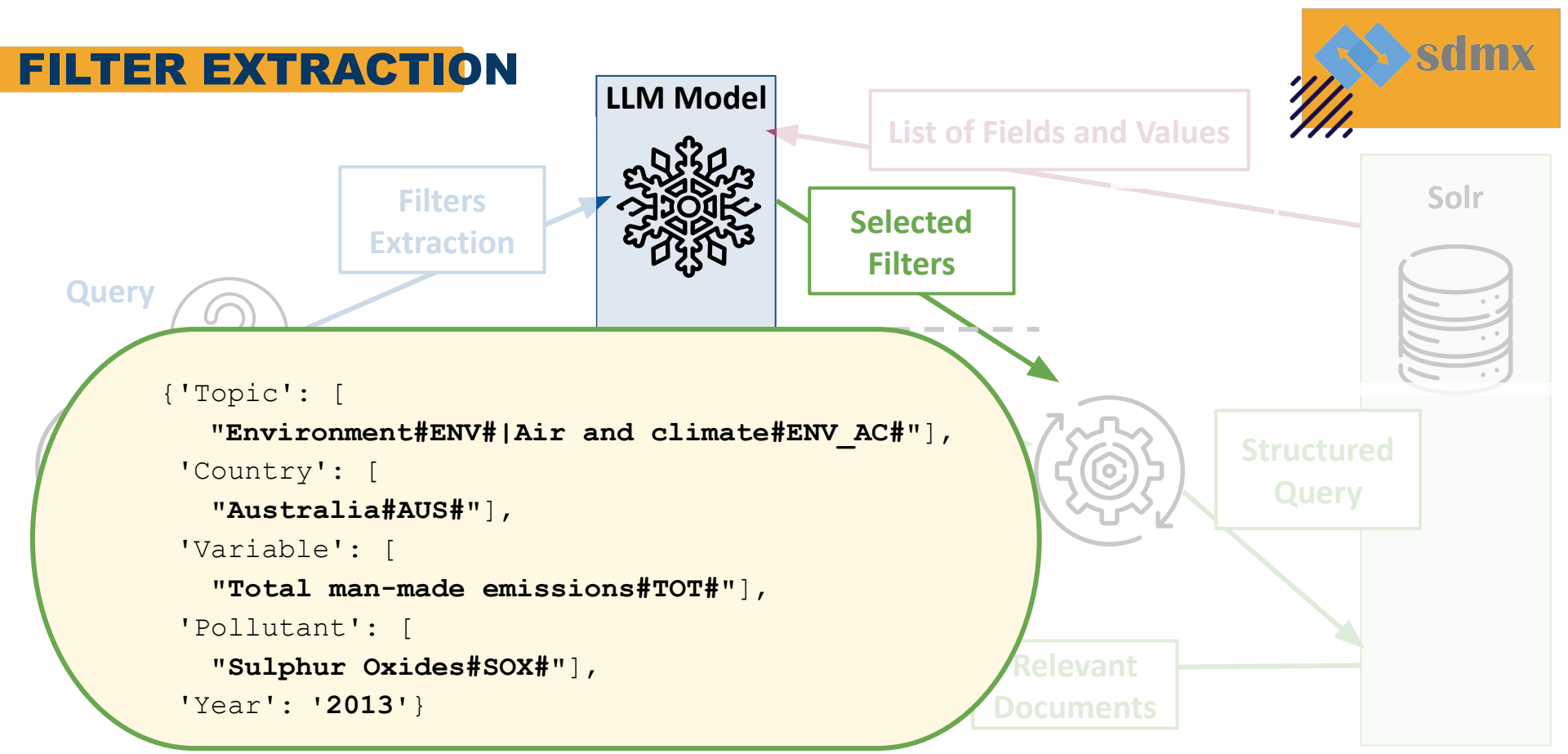
PROMPT

1) **OBJECT REPRESENTATION**
Provide input data (e.g. JSON representation) to the model

1) **QUERY PARSING**
Request a similar representation (i.e. subset) based on the input query

1) **FORMAL REQUIREMENTS**
Specify how the output should be formatted, including any constraints or specific criteria to be met.

FILTER EXTRACTION



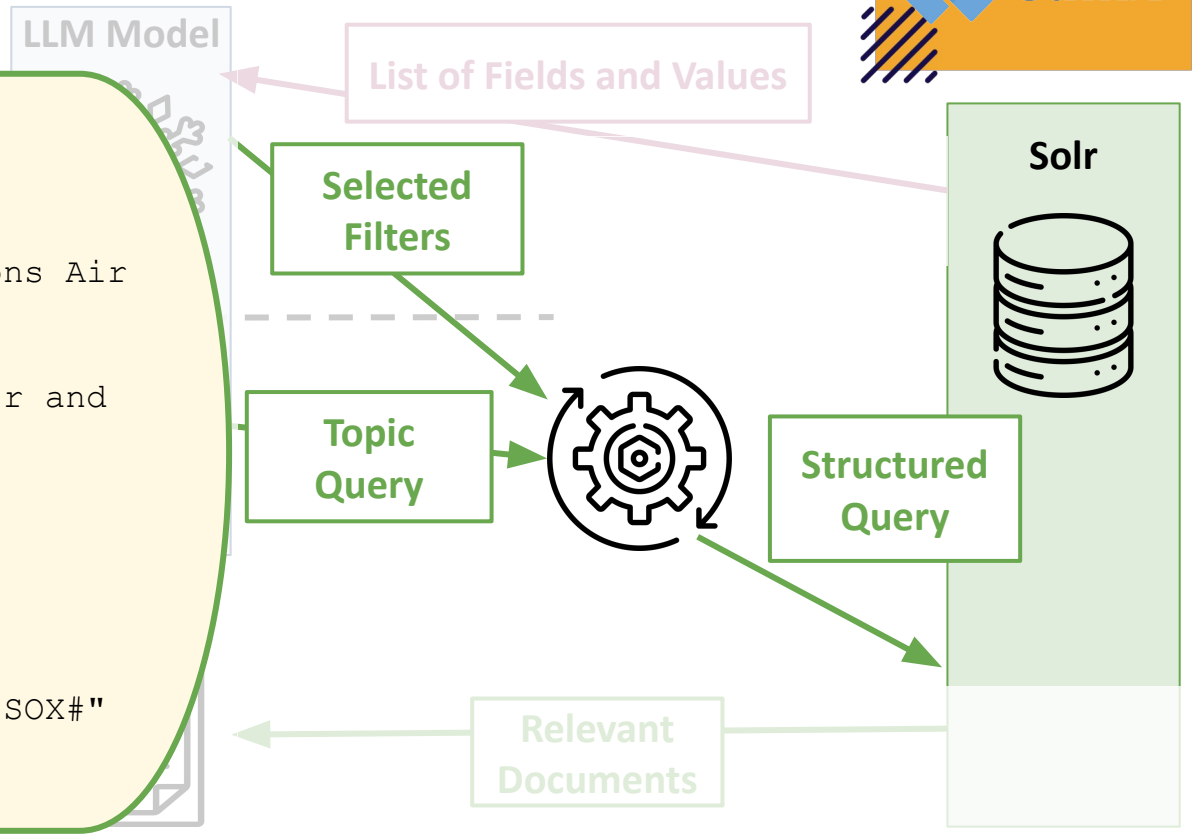

```
{'Topic': [
   "Environment#ENV#|Air and climate#ENV_AC#"],
 'Country': [
   "Australia#AUS#"],
 'Variable': [
   "Total man-made emissions#TOT#"],
 'Pollutant': [
   "Sulphur Oxides#SOX#"],
 'Year': '2013'}
```

What were the sulfur oxide emissions in Australia in 2013?

STRUCTURED QUERY

SOLR QUERY

```
q=
title:(Sulfur dioxide emissions Air
... Acid rain)
OR topic:"Environment#ENV#|Air and
climate#ENV_AC#"
OR country:"Australia#AUS#"
OR variable:"Total man-made
emissions#TOT#"
OR Pollutant:"Sulphur Oxides#SOX#"
OR 'Year': '2013'
```



DOC RETRIEVAL

LLM Model

SEARCH RESULTS

```
"response":{
  "numFound":1,
  "start":0,
  "numFoundExact":true,
  "docs":[{
    "Title":"Emissions of air pollutants",
    "Dimension":["Country", "Pollutant", "Variable", "Year"]
  }]
}
```

Query

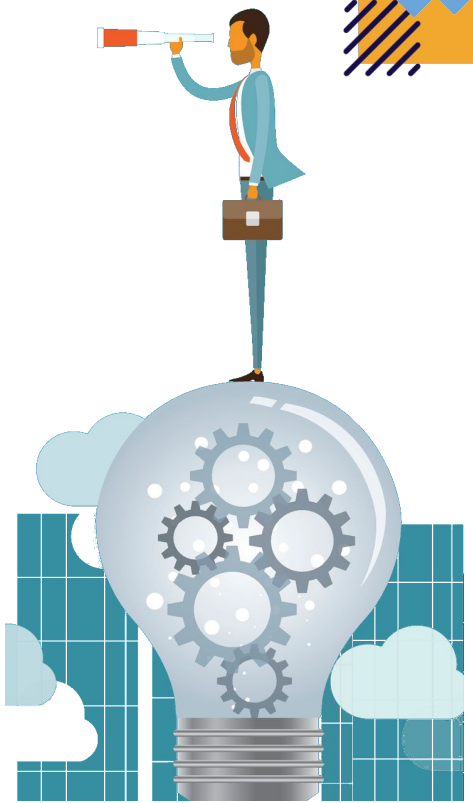
Solr

Answer

Relevant Documents

AGENDA

- Use Case Overview
- From Natural Language to Structured Queries
- Findings
- The Road to Production



MODEL CONSIDERATIONS

- **[Model Selection]:** https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
- **[Rationale for Current Choice]:** No deep evaluations or comparisons with alternative models happened in the POC
- **[Future Works]:**
  - **Explore, analyze and compare generalist models**
  - Potentially undertake **our own fine-tuning** for the specific extractive task

PROMISING ASPECTS

- **Overcome** the **lexical matching**

  land of kangaroos → [Country] AUSTRALIA

  tobacco consumption → [Topic] SMOKING/RISK FACTORS FOR HEALTH

PROMISING ASPECTS

- **Explainability** for **selected filters**

Analyze input text: *"cost per square meter for family houses in italy"*

cost per square meter → *pricing or valuation* → 'Priced unit' or 'Value'

family houses → *type of property* → 'Real estate type'

italy → *location* → 'Reference area' or 'Borrowers' country'

We need to identify which dimensions and their corresponding values are most relevant to the input text **"cost per square meter for family houses in italy"**. To do this, we will look for dimensions that are directly related to real estate, housing, or geographic location, specifically within Italy.
…

PROMISING ASPECTS

- **Explainability** for **selected filter**

Analyze input text: *"cost per squar*

cost per square meter → *pricing o*

family houses → *type of p*

italy → *location*

IDEA!

Integrate as an "Assistant" feature

to guide users in choosing the most suitable filters

PROMISING ASPECTS

- **Promising potential in early results (POCs):**
  - good **results** (using a commercial out-of-the-box model!)
  - **straightforward** implementation for such a challenging and **complex task**
  - model's **adaptability** to the context

LIMITATIONS

1 FUNCTIONAL

LLM weaknesses in the
language/query semantic comprehension

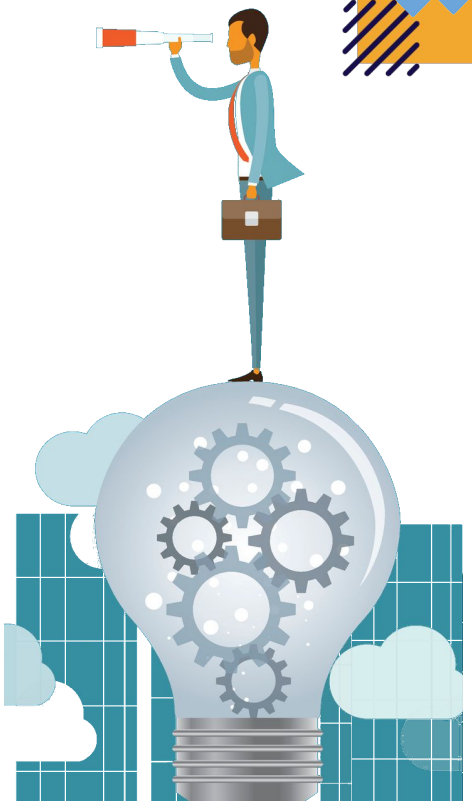
2 FORMAL

LLM weaknesses in complying with:

- the **problem definition**
- the required **output format**

AGENDA

- Use Case Overview
- From Natural Language to Structured Queries
- Findings
- The Road to Production



THE ROAD TO PRODUCTION

- [UX] **Design the user experience**
  - Filtering assistance?
  - Transparent query parsing?
- [LLM] **Select the best model to date**
  - Can we fine-tune promising models specifically for the task?
- [LLM] **Refine the prompts according to the model**
  - Can we reduce functional and formal errors?

THE ROAD TO PRODUCTION

- [LLM] **Implement integration tests with the most common failures** → LLM/prompt engineering to solve them
- [LLM] **Study additional libraries** to make the prompt more “programmed” and “automatically tuned” and less “trial-and-error”
  - Highly depend on the LLM available
- [Performance] **Stress test** the solution
- [Quality] **Set up queries/expected documents**

THANK YOU!

SCAN ME

@seaseltd @sease-ltd @seaseltd @sease_ltd