Search Quality Evaluation in the Era of Large Language Models: Dataset Generator

Speaker: Alessandro Benedetti

Search Solutions - 26th November 2025

Sease | Information Retrieval Applied

Who are we

ALESSANDRO BENEDETTI

Director + R&D Software Engineer @ Sease

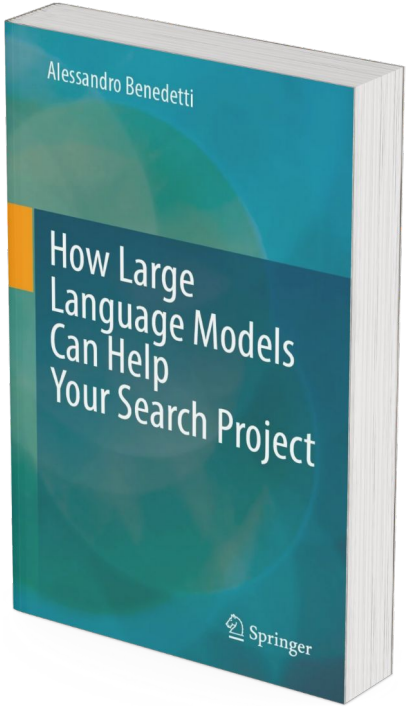
- Born in **Tarquinia** (ancient Etruscan city in Italy)
- Master Degree in **Computer Science**
- Program Committee member for **ECIR, SIGIR** and **Desires**
- **Apache Solr Chair of the PMC + Apache Lucene/Solr committer**
- Elasticsearch/OpenSearch expert
- Semantic search, NLP, Machine Learning technologies passionate
- Beach Volleyball player and Snowboarder

- 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

How Large Language Models Can Help Your Search Project



COMING SOON

London Search Week

NOVEMBER 24-27

FIVE DAYS TO LEARN, BUILD, AND BENCHMARK NEXT-GEN SEARCH ACROSS THE STACK.

MON 24TH NOVEMBER

2-5 PM - WE ARE SEARCH PRODUCT MANAGERS BARCAMP
6-9 PM - LONDON INFORMATION RETRIEVAL & AI MEETUP

TUE 25TH NOVEMBER

9 AM-6 PM - SEARCH SOLUTIONS TUTORIAL DAY

WED 26TH NOVEMBER

9:30 AM-9 PM - SEARCH SOLUTIONS CONFERENCE DAY

THU 27TH NOVEMBER

6-9 PM - AGENTIC AI MEETUP - SEARCH HAS A NEW (AGENTIC) USER: WHAT NOW?

FRI 28TH NOVEMBER

9 AM-6PM - HACKDAY LUCENE/SOLR CONTRIBUTIONS

Overview

1. Search Quality Evaluation Problem
2. The Dataset Generator Introduction
3. How the Dataset Generator works
4. Next Steps

Overview

1. Search Quality Evaluation Problem
2. The Dataset Generator Introduction
3. How the Dataset Generator works
4. Next Steps

Why measure search quality?

- You **can't improve** what you **can't measure**
- It opens a tangible **comparison between more query designs**
- It **reveals** dataset and model **weaknesses**
- It's the foundation for **long-term iteration**
- It unlocks **continuous improvement** through measurement

Image by juicy_fish on Freepik

Implicit and Explicit feedback

Implicit

- Feedback inferred **automatically** from user behavior.
- Clicks on search results, dwell time, scroll depth, query reformulations, skipped results.

- Easy to gather at scale.
- Reflects real-world behavior.
- Noisy and ambiguous (e.g., click ≠ satisfaction)

Explicit

- Feedback directly provided by users in a **deliberate** way.
- Ratings (e.g., 1–5 stars), relevance judgments, thumbs up/down, written comments, likes.

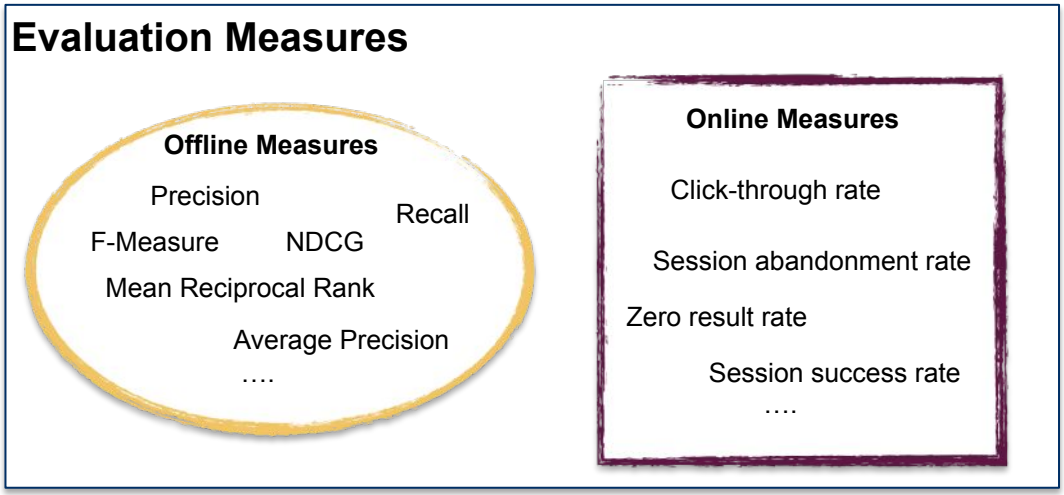
- High precision and interpretability.
- Expensive and time-consuming.
- Hard to scale and may suffer from subjectivity.

Evaluation Measures

Evaluation measures for an information retrieval system formalise how well a search system satisfies its user information needs.

Measures are generally split into two categories: **online** and **offline measures**.

We'll focus on **offline measures.**



Offline measures

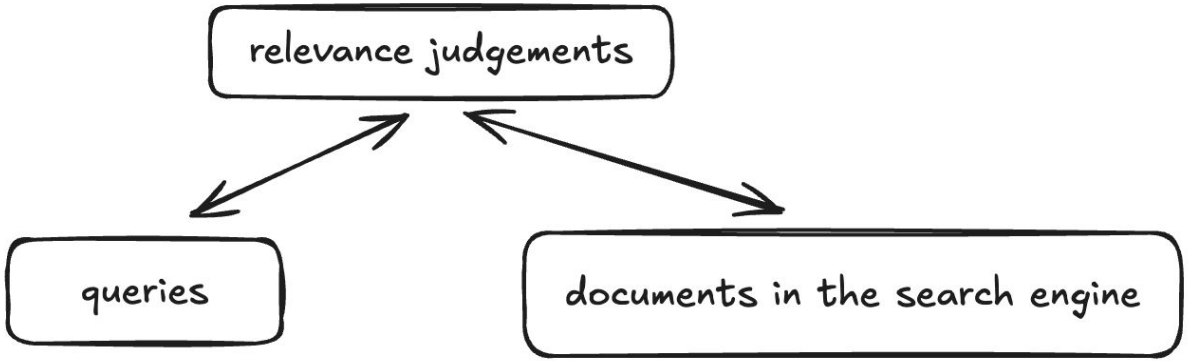
Advantages:

- Find **anomalies in data**, like: weird distribution of the features, strange collected values, ...
- Check how the **ranking function performs** before using it in production: implement improvements, fix bugs, tune parameters, ...
- Save **time** and **money**. Put in production a bad ranking function can worse the user experience on the website.

Offline search quality evaluation data

In the scenario of offline search quality evaluation, the main obstacle is the lack of **queries** and **relevance judgements**.

Without labeled **query–document pairs**, it becomes difficult to assess retrieval quality.



The Challenge of Obtaining Labeled Data

- **Expensive & time-consuming:** Requires expert annotators.
- **Accuracy:** Labels become less accurate due to the fact that labeling is a boring job and requires focus for a long period of time.
- **Subjective:** Relevance may depend on context, user intent, or domain.
- **Scalability issue:** Hard to label enough data for large corpora or many queries.

Image by vectorjuice on Freepik

Why LLMs change the game for synthetic dataset generation

Generated by ChatGPT

Overview

1. Search Quality Evaluation Problem
2. The Dataset Generator Introduction
3. How the Dataset Generator works
4. Next Steps

Dataset Generator

Tutorial material

rated-ranking-evaluator repo:

llm-search-quality-evaluation repo:

llm-search-quality-evaluation-tutorial repo:

The team behind the Dataset Generator

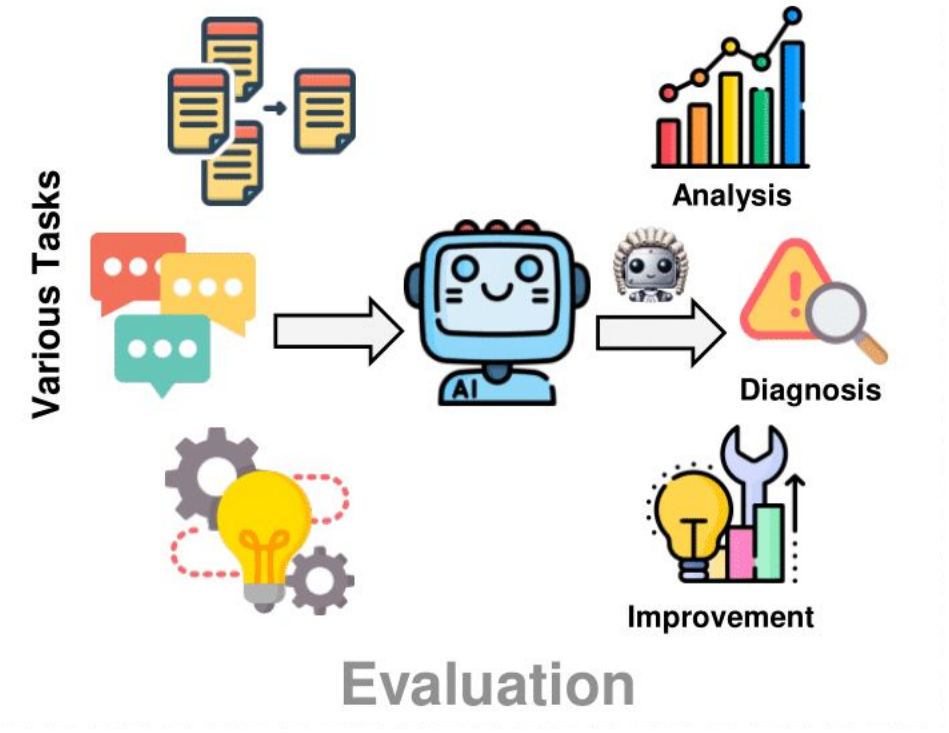
DANIELE ANTUZI
R&D SOFTWARE ENGINEER
SEARCH CONSULTANT

NICOLÒ RINALDI
SOFTWARE ENGINEER/DATA SCIENTIST

NAZERKE SEIDAN
IR/ML
SOFTWARE ENGINEER

From ai4business.it

LLM-as-a judge for evaluation



Assumption: you have only the documents indexed

Generate queries

Generate relevance labels (with rationale)

Functionalities

Supported search engines (input)

elasticsearch

Main Inputs

- Documents (quantity, filters)
- (optional) existing Queries
- LLM configurations
- Search engine connection parameters (URL, collections...)

Interpretability

- Possibility to enable LLM “explainability”
- Possibility to enrich the set of document scored for each query using a query template and retrieve more documents directly from your search engine

Dataset Generator Output formats

quepid	mteb	RRE

quepid

```
query,docid,rating
What is the nature and purpose of the mission the helicopter was on?,d1,2
How did the helicopter crash in the Colombian jungle?,d1,2
What is the involvement of U.S. technology or personnel in the Colombian drug war?,d1,2
Why did the police use a Taser on the python?,d2,2
What is the nature and purpose of the mission the helicopter was on?,d2,0
How did the helicopter crash in the Colombian jungle?,d2,0
What is the involvement of U.S. technology or personnel in the Colombian drug war?,d2,0
Why did the police use a Taser on the python?,d1,0
```

mteb

```
{"id": "d1", "title": "Helicopter Crashes in Colombian Drug War, Kills 20", "text": "BOGOTA, … army said."}
{"id": "d2", "title": "Police Use Taser on Python to Free Man", "text": "UNIONTOWN, Pa. … not let go."}
{"id": "d3", "title": "Zidane apologizes for head butt", "text": "French soccer … and sister."}
{"id": "d4", "title": "Iraqi PM Says Sticking to January Election Plan", "text": "Iraqi Prime … insurgent."}
{"id": "d5", "title": "Ga. Crematory Operator to Plead Guilty", "text": "Relatives … guilty plea."}
{"id": "d6", "title": "Long-driving Kuehne finally breaks out of 2004 slump", "text": "Hank Kuehne … Kuehne"}
```

```
{"id": "q1", "text": "What is the nature and purpose of the mission the helicopter was on?"}
{"id": "q2", "text": "How did the helicopter crash in the Colombian jungle?"}
{"id": "q3", "text": "What is the involvement of U.S. technology or personnel in the Colombian drug war?"}
{"id": "q4", "text": "Why did the police use a Taser on the python?"}
```

```
{"query_id": "q1", "doc_id": "d1", "rating": 2}
{"query_id": "q2", "doc_id": "d1", "rating": 2}
{"query_id": "q3", "doc_id": "d1", "rating": 2}
{"query_id": "q4", "doc_id": "d2", "rating": 2}
{"query_id": "q1", "doc_id": "d2", "rating": 0}
{"query_id": "q2", "doc_id": "d2", "rating": 0}
```

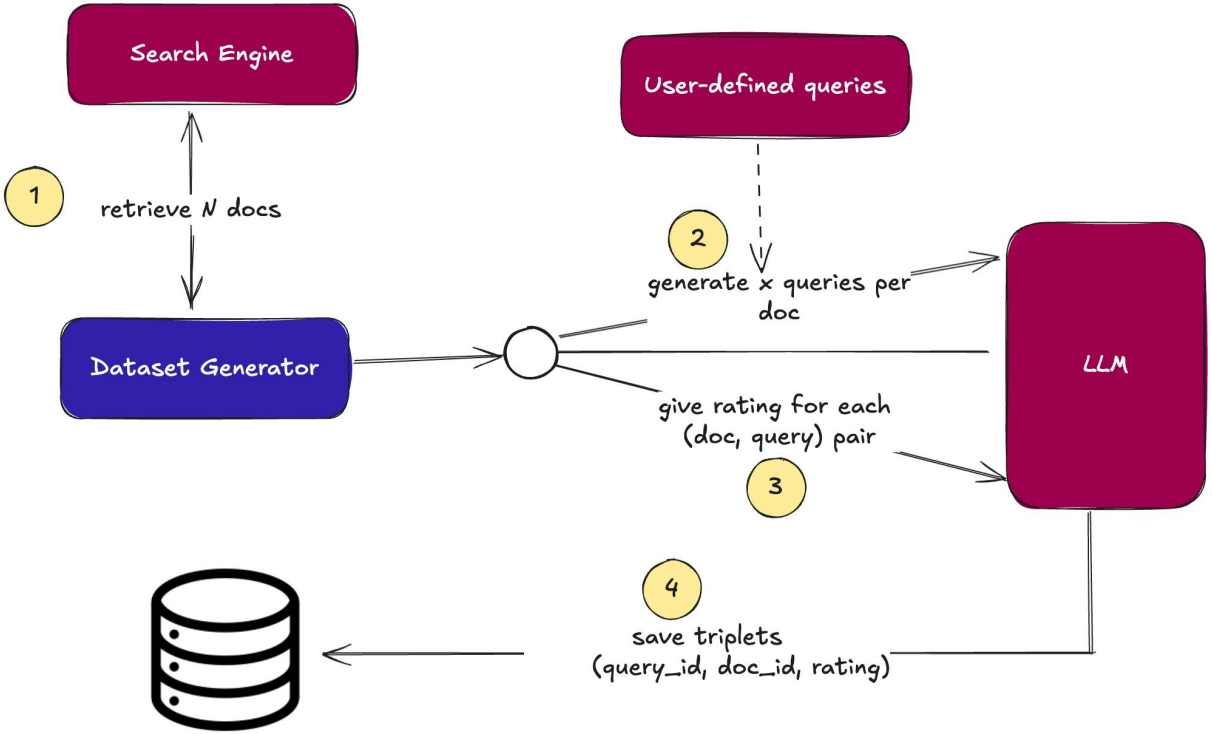
RRE

```
{
  "index": "testcore",
  "id_field": "id",
  "query_placeholder": "$query",
  "query_groups": [
    {
      "name": "What is the nature and purpose of the mission the helicopter was on?",
      "queries": [
        {
          "template": "template_solr.json",
          "placeholders": {
            "$query": "What is the nature and purpose of the mission the helicopter was on?"
          }
        }
      ],
      "relevant_documents": {
        "2": ["d1"],
        "1": ["d2", "d32", "d107", "d13"],
        "0": ["d175", "d189", "d162", "d88", "d126", "d65", "d184" ]
      }
    }
  ]
}
```

Overview

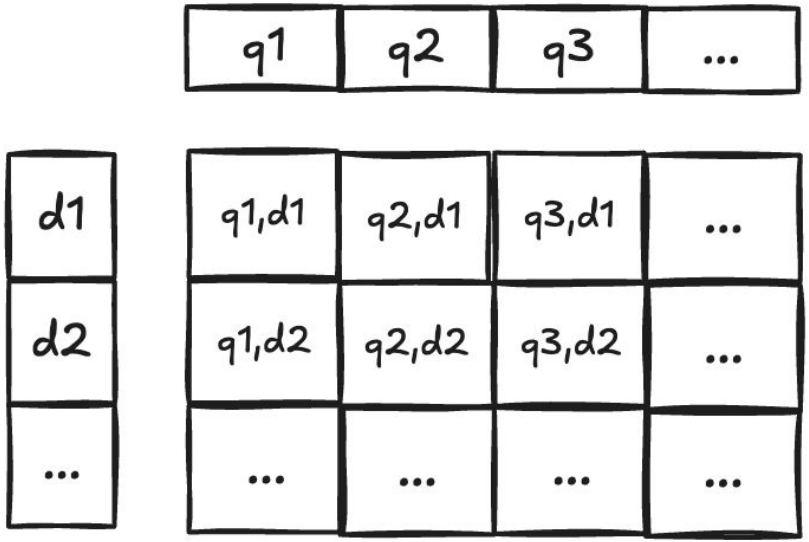
1. Search Quality Evaluation Problem
2. The Dataset Generator Introduction
3. How the Dataset Generator works
4. Next Steps

First step: query generation



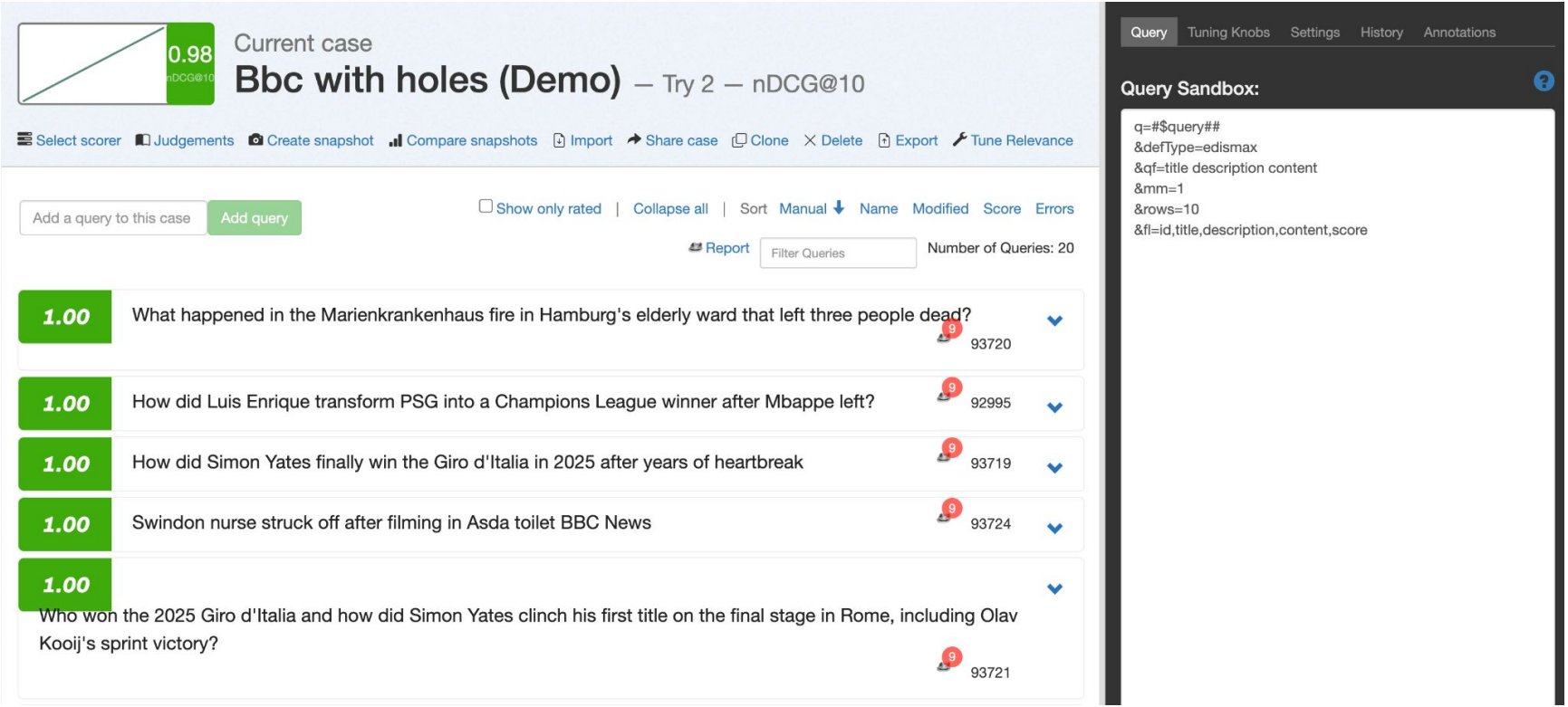
About cartesian product

3



<u>Recommendation</u>: disable this if not strictly needed, since it increases consistently the number of LLM calls!

Second step: retrieval with query template



Only first one is rated!

1.00 What happened in the Marienkrankenhaus fire in Hamburg's elderly ward that left three people dead?

9 93720

Score All

-- Toggle Notes | Explain Query | Missing Documents | Set Options | Move Query | Delete Query

2

Three dead after fire in elderly ward at Hamburg hospital - BBC News

title: Three dead after fire in elderly ward at Hamburg hospital - BBC News

content: Three dead after fire in elderly ward at Hamburg hospital The fire at the Marienkrankenhaus in Hamburg was reported shortly after midnight Three people have died after a fire broke out overnight at a hospital ward caring for elderly people in Hamburg. The north German city's fire brigade said they were unable to reach the three victims during the rescue effort. They died at the scene. The fire at the Marienkrankenhaus was reported shortly after midnight. Around 220 emergency workers were dispatched to tackle the blaze and treat patients. More than 35 people were injured. Two people were taken to another hospital in the city for treatment. One person has life-threatening injuries, said the fire brigade. They added 18 others had suffered serious injuries and another 15 people had minor wounds. Some of the injured were treated in the Marienkrankenhaus's emergency room by hospital staff. When the first firefighters arrived, a ground floor room at the hospital was already ablaze and the fire was threatening to spread to the room above.

id: 948f5374-36f7-46f9-b13a-b57f66bb1fd2

Rank: #1

Matches

weight(content:marienkrankenhaus in 10) [SchemaSimilarity], result of:

weight(title:ward in 10) [SchemaSimilarity], result of:

weight(title:elderly in 10) [SchemaSimilarity], result of:

Show 6 More

--

'Dad tried to kill us': The fire that devastated Australia - BBC News

title: 'Dad tried to kill us': The fire that devastated Australia - BBC News

content: 'Dad tried to kill us': The fire that devastated Australia Filicide - when a parent intentionally kills their child - is the second most common form of domestic homicide in Australia The night comes back in fragments: the sound of exploding glass, the frantic call made to police, the tiny shivering bodies emerging from the flames. Eve's hand shakes as she pieces it together. She is sitting in her living room in Western Sydney, the burnt-out shell of her neighbour's house - now a crime scene - visible through the blinds. What happened on this quiet street in the early hours of Sunday morning is hard to reconcile. A fire that would leave three children dead, including a five-month-old girl, and four more hospitalised alongside their mother. And a stunning allegation: that

Matches

weight(title:that in 15098) [SchemaSimilarity], result of:

weight(content:fire in 15098) [SchemaSimilarity], result of:

weight(description:three in 15098)

Why using a template is important

- Documents in input may be a subset of entire corpus
- Generated queries may retrieve un-rated documents
- You can run the queries and use the LLM to fill the gaps

In this way, you can also visualize your results with Quepid for example.

After using the template: all retrieved docs are rated!

0.99 nDCG@10

Current case

Bbc (Demo) — Try 2 — nDCG@10

Select scorer | Judgements | Create snapshot | Compare snapshots | Import | Share case | Clone | Delete | Export | Tune Relevance

Add a query to this case | Add query

Show only rated | Collapse all | Sort Manual Name Modified Score Errors

Report | Filter Queries | Number of Queries: 30

0.93	How many migrants crossed the English Channel in a single day in 2025, according to BBC News?	93724
1.00	What was Ukraine's 'Spider's Web' drone attack against Russia and how many drones were used?	93489
0.95	What is daily life like for Ukrainians living under Russian occupation in Melitopol and other occupied cities, including acts of resistance and pressures like passport demands, according to BBC News.	93724
1.00	Briton indicted for allegedly exporting US military technology to China; arrested in Serbia and facing US extradition	93719
1.00	What were the criminal gang links of Eddie Lyons Jr and Ross Monaghan, the two Scots shot dead in Fuengirola, Spain?	93721

Query | Tuning Knobs | Settings | History | Annotations

Query Sandbox:

```
q=#$query##
&defType=edismax
&qf=title description content
&mm=1
&rows=10
&fl=id,title,description,content,score
```

Configuration file

```
query_template: "templates/template_solr.json"        <- be careful: this file must follow RRE guidelines
search_engine_type: "solr"                        <- support: “solr”, “vespa”, “opensearch”, “elasticsearch”
collection_name: "testcore"
search_engine_url: "http://localhost:8983/solr/"      <- instance MUST be up and running
documents_filter:                                     # if any
  - genre:
      - "horror"
number_of_docs: 100
doc_fields:
  - "title"
  - "description"
queries: "queries.txt"                                # if any
generate_queries_from_documents: true                 <- if not set: generation enabled
num_queries_needed: 10
relevance_scale: "graded"                             <- graded ({0, 1, 2}), binary ({0, 1})
llm_configuration_file: "configs/dataset_generator/llm_config.yaml"
max_query_terms: 5                                    <- if not set: let LLM decide
output_format: "quepid"                               <- “rre”, “quepid”, “mteb”
output_destination: "resources"
save_llm_explanation: true
llm_explanation_destination: "resources/rating_explanation.json"
datastore_autosave_every_n_updates: 50                <- if not set: autosave disabled
enable_cartesian_product: false                       <- if not set: cartesian product enabled
```

Query template file structure

```
"queries": [
  {
    "template": "only_q.json",
    "placeholders": {
      "$query": "fender"
    }
  },
  {
```

```
{
  "template": "filter_by_language.json",
  "placeholders": {
    "$query": "Fender",
    "$lang": "eng"
}
```

only_q.json

```
{
  "q" : "$query"
}
```

filter_by_language.json

```
{
  "q" : "$query",
  "fq" : "language:$lang"
}
```

Example of llm_config.yml file

```yaml
# OpenAI LLM
name: openai            # supported: openai, gemini

# Chat model name
model: gpt-5-nano-2025-08-07

# Maximum number of tokens the model may return
max_tokens: 2000

# Environment variable where LLM API key is stored
api_key_env: OPENAI_API_KEY
```

How to run RRE Dataset Generator

```
Step 1: ensure that your search engine is up and running

Step 2: clone rated-ranking-evaluator repo

Step 3: cd rre-tools

Step 4: prepare configuration file

Step 5: uv run dataset_generator

That's it!
```

Results: explanation

Doc: Ukrainian forces outnumbered Russians by eight to one in last week's counter-attack in the Kharkiv region, Russia's top occupation official there says.
Vitaly Ganchev told Russian TV that Ukraine's army had taken villages in the north and broken through to the Russian border.
Ukraine says it has regained control over 3,000 sq km (1,158 sq miles) of territory in a potential breakthrough in the six-month war.
The BBC cannot verify these figures.
The Ukrainian army says it took back 20 villages in the past 24 hours alone, in its continued counter-offensive in the north-east of the country.
It also said its forces have taken control of around 500 sq km in the southern Kherson region of the country.
UK defence officials say the Ukrainian army's successes will have "significant implications" for Russia's overall operational design.
Kremlin spokesman Dmitry Peskov was, however, undeterred, saying operations in Ukraine would continue "until all the tasks that were initially set" had been fulfilled.

Query: Life under occupation in Ukraine

Explanation: The article is war news about counter-attacks and territorial changes and includes a mention of an occupation official in the Kherson region, which signals an occupation context. However, it does not discuss daily life under occupation, civilian experiences, restrictions, or living conditions. Because it touches on occupation only superficially and not the core topic of “life under occupation,” it is only marginally relevant (score 1) to the query.

Overview

1. Search Quality Evaluation Problem
2. The Dataset Generator Introduction
3. How the Dataset Generator works
4. Next Steps

Next steps

- Documents cluster for query generation
- Use input queries as few shots for the LLM
- Use some existing ratings as few shots for the LLM
- New tools (Vector Search Doctor!)

Vector Search Doctor

1. Did I pick the right embedding model?

2. Did I pick the right vector search implementation?

Sease | Information Retrieval Applied

THANK YOU

sease.io