Swiss Re

IDS Conference

# GenAI: Developing and deploying a specialized underwriting AI assistant

June 2024

Louis DOUGE

# Life Guide

Life Guide is a L&H underwriting resource, helping clients understand current and future risks, and translates them into ratings to build strong and sustainable portfolios.

Industry's #1 L&H underwriting resource

Trusted by underwriters in over 100 countries

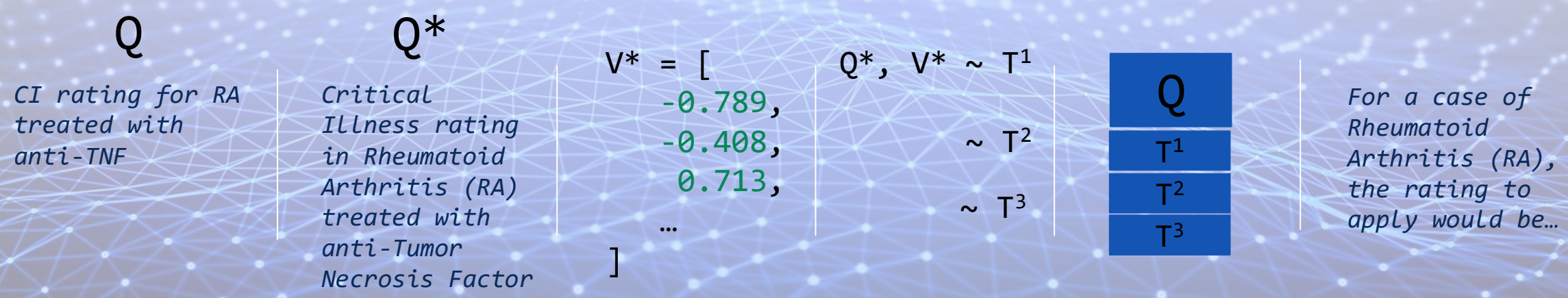
Receives over 23 million hits annually

**Goal of creating a specialized assistant**

- Support underwriters with a human like assistant
- Enable interaction with a bot in natural language
- Ask a single question and receive a targeted and concise answer

# Retrieval-Augmented Generation (RAG) pipeline

- RAG introduces an information retrieval component to fetch relevant information from *external data*
- The user query and the relevant information are used by the LLM to produce an answer



# Constructing the *external data* store: chunking strategy and vector database

**Life Guide Manual**

**Text chunks**

$T^1$

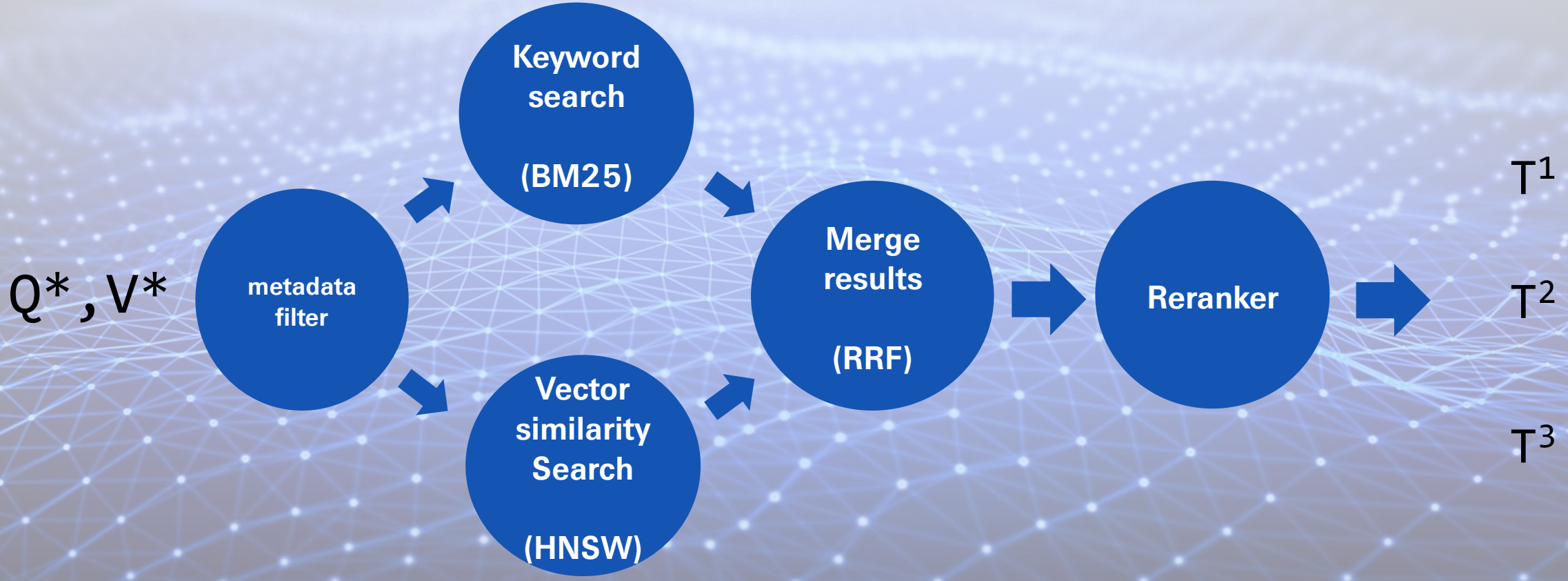
$T^2$

$T^3$

**Vector Database**

```
{
    "id": "asdf-64-651asdf",
    "text": "Asthma \n Prevalence of disease...",
    "embedding": [.53,.63,..., .23],
    "metadata1": "adsfj",
    "metadata2": "dasff"
}
{
    "id": "qwe-12-yxc",
    "text": "Asthma \n Symptoms include...",
    "embedding": [.73,.68,..., .72],
    "metadata1": "xcbvcxyb",
    "metadata2": "ezrtre"
}
{
    "id": "<yxyc-95-sdfd",
    "text": "Asthma \n Diagnosis is...",
    "embedding": [.35,.879,..., .45],
    "metadata1": "qfesdv",
    "metadata2": "dsfzifgf"
}
```

# Retrieving the relevant information



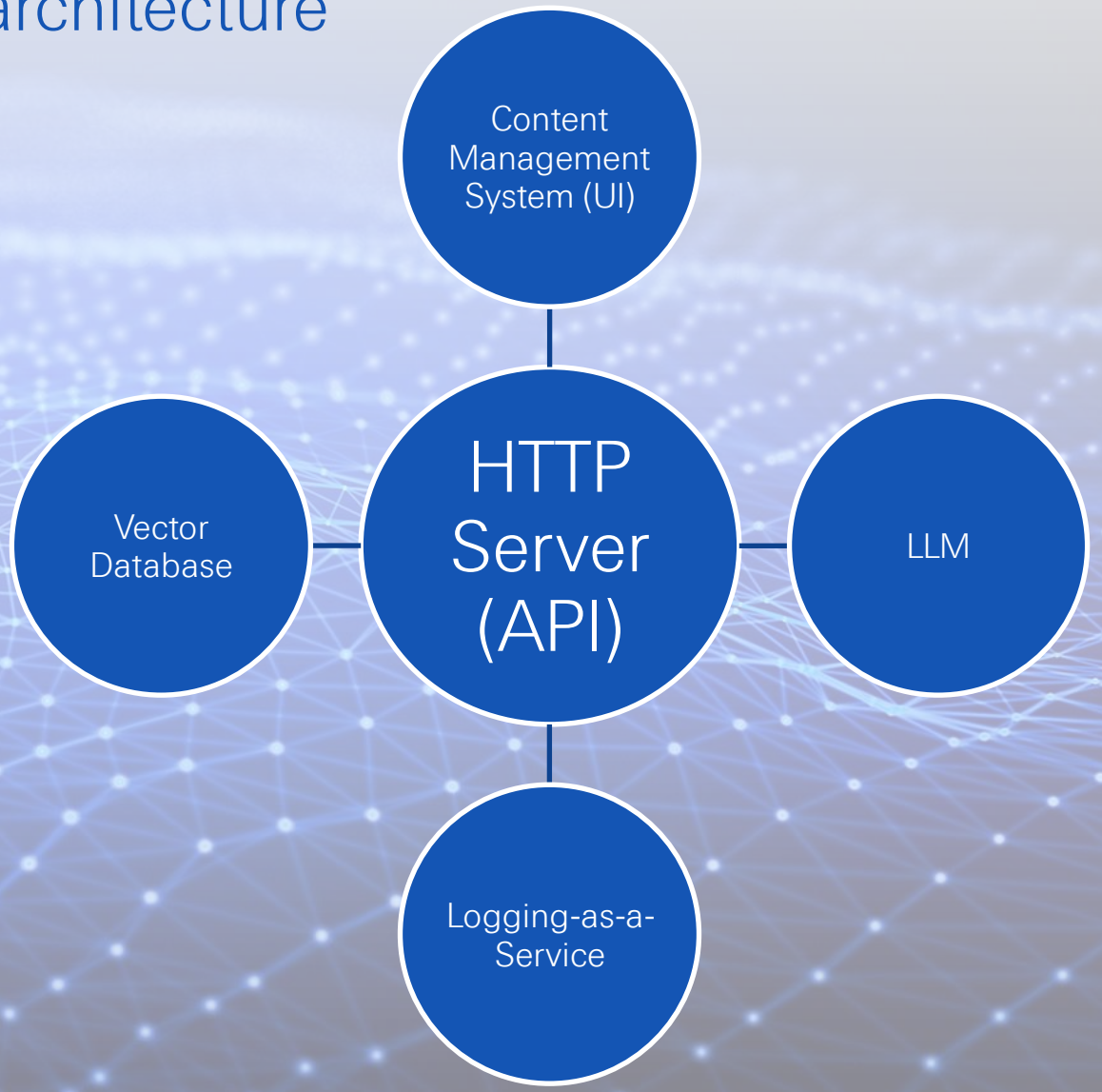

BM25 : TF-IDF-like relevance ranking algorithm
HNSW : Hierarchical Navigable Small Worlds
RRF : Reciprocal Rank Fusion

# Engineering of the deployment: architecture

**Microservices architecture**

**Importance of latency optimisation**

**Asynchronous logging**



# Engineering of the deployment: monitoring with Kibana

**Number of session per user**

**Latency of each RAG steps**

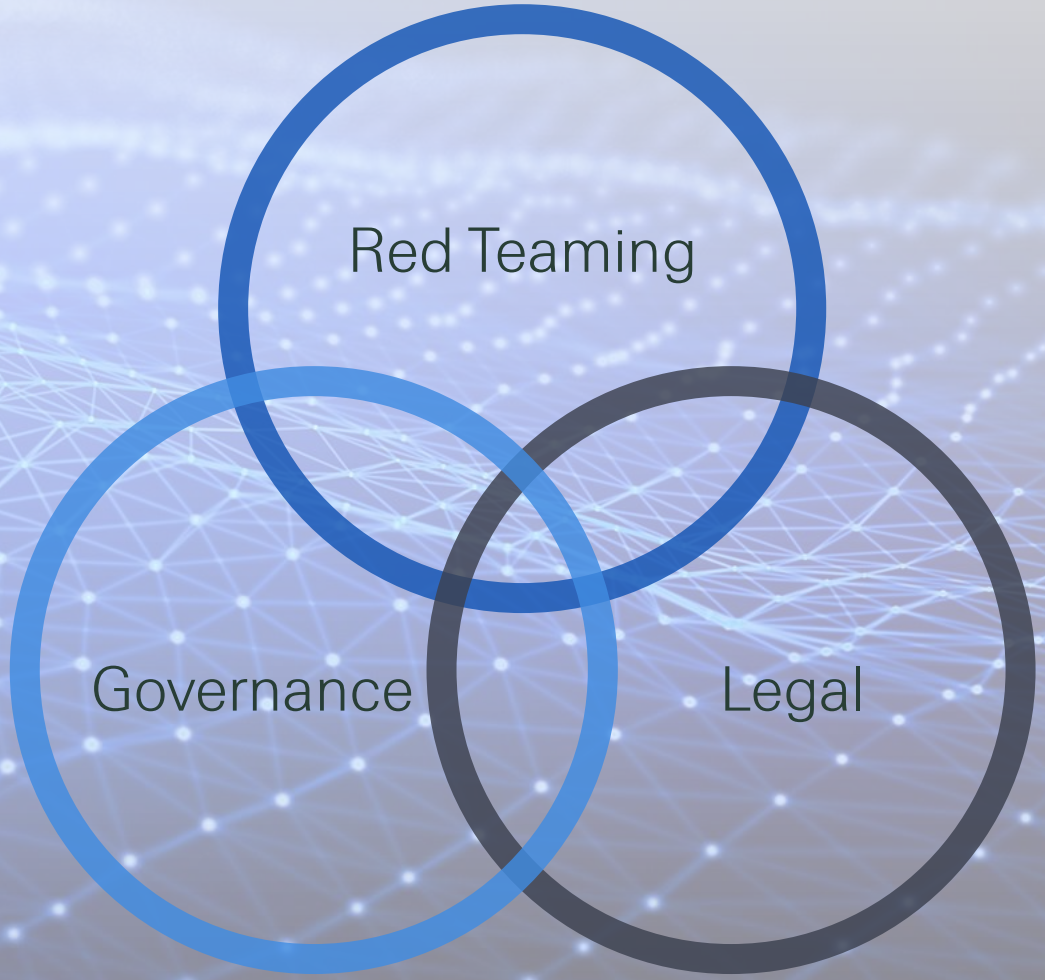
**Token utilization**

**Errors**

**Running costs**

**...**

# Safety through testing

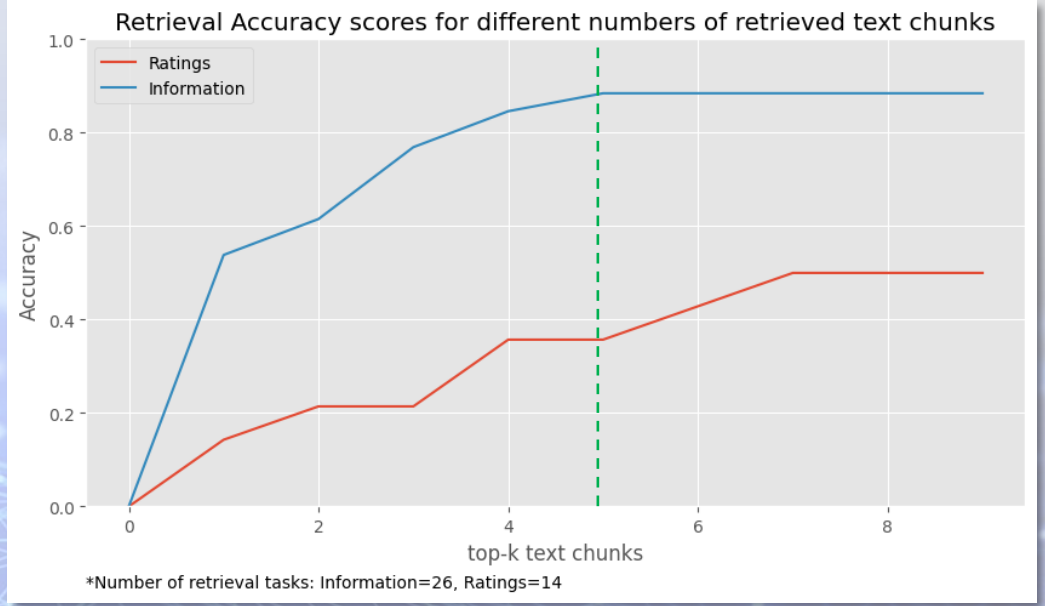


# Programmatic evaluation

**Evaluation dataset**
- Synthetically generated
- Reviewed & Validated by experts

**Metrics**
- Individual pipeline components metrics
- End-to-end evaluation:
  - RAGAs methodology (*)
  - No annotated samples required

**Importance of a fast evaluation turnover**




**RAGA's *Context relevance* metric prompt**

*Please extract relevant sentences from the provided context that can potentially help answer the following question. If no relevant sentences are found, or if you believe the question cannot be answered from the given context, return the phrase "Insufficient Information". While extracting candidate sentences you're not allowed to make any changes to sentences from given context*

(*) Shahul et al. *RAGAs: Automated Evaluation of Retrieval Augmented Generation.* arXiv:2309.15217v1, 2023

# Human evaluation

**Survey**
- Understand overall impression of product
- Gauging adaptability and trust

**User Feedback**
- Direct collection of model performance and its accuracy
- Qualitative feedback on what needs improving

Swiss Re

# Thank you!

## Contact us

Louis Douge
Senior Data Scientist
Louis_Douge@swissre.com

## Follow us

# Legal notice



The information and opinions contained in the presentation are provided as at the date of the presentation and may change. Although the information used was taken from reliable sources, Swiss Re does not accept any responsibility for its accuracy or comprehensiveness or its updating. All liability for the accuracy and completeness of the information or for any damage or loss resulting from its use is expressly excluded.