



# Roundtable Introduction

E. Lancon

08/13/26

    @BrookhavenLab

# Today

- Update on SciBot
- User experience of repeating a STAR analysis - **Ankush**
- Evolution of Datasheet/Datacard concept - **Maxim**

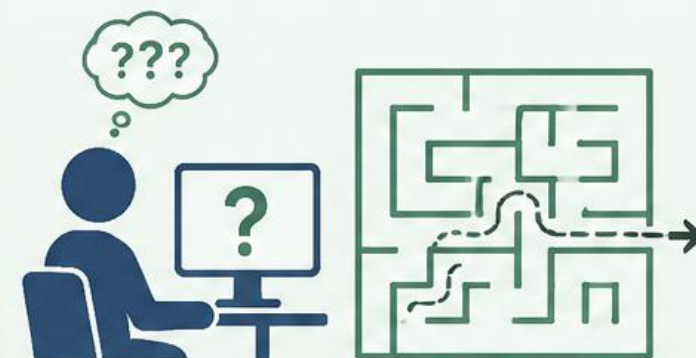
# User Experience of Repeating a STAR Analysis

A reproduction attempt **six years later** by someone outside the original analysis team

The goal is not simply to verify that the analysis can be reproduced, but to:

- Identify the information, documentation, and tools needed by someone unfamiliar with the original analysis
- Capture lessons learned for documenting future analyses
- Identify the skills an AI agent would need to reproduce an analysis and adapt as the software and computing environment evolve
- Understand how the process can scale to the **hundreds of analyses** that may need to be preserved

## Reproduction Attempt Today By a New User



-  Finding the right information
-  Understanding the environment and dependencies
-  Getting code, data, and workflows to run
-  Validating intermediate steps and results
-  Missing context and tacit knowledge

# Datasheets / Datacards

Help accessing the data

But having access to the data is only the beginning.

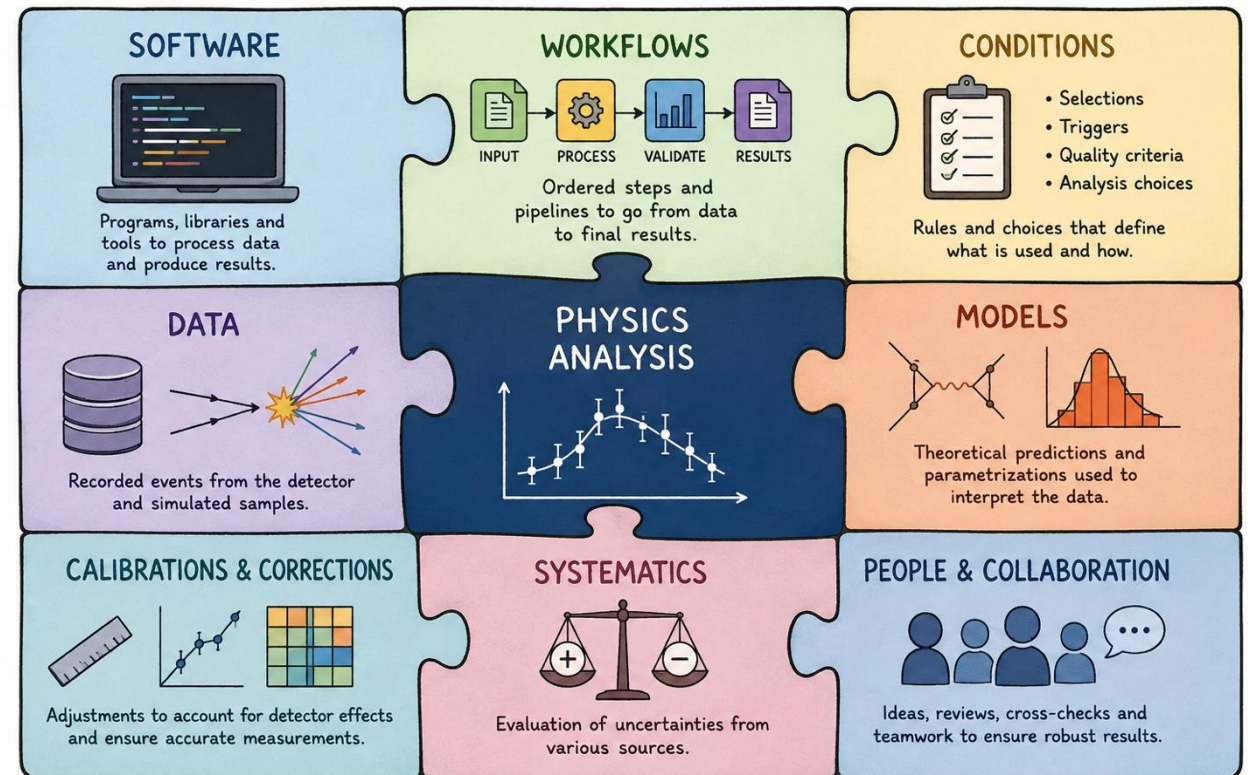
It does not tell us how to use the data to perform a physics analysis.

**A physics analysis is a complex assembly of interconnected objects:**

data • software • workflows • conditions •  
calibrations • corrections • systematics

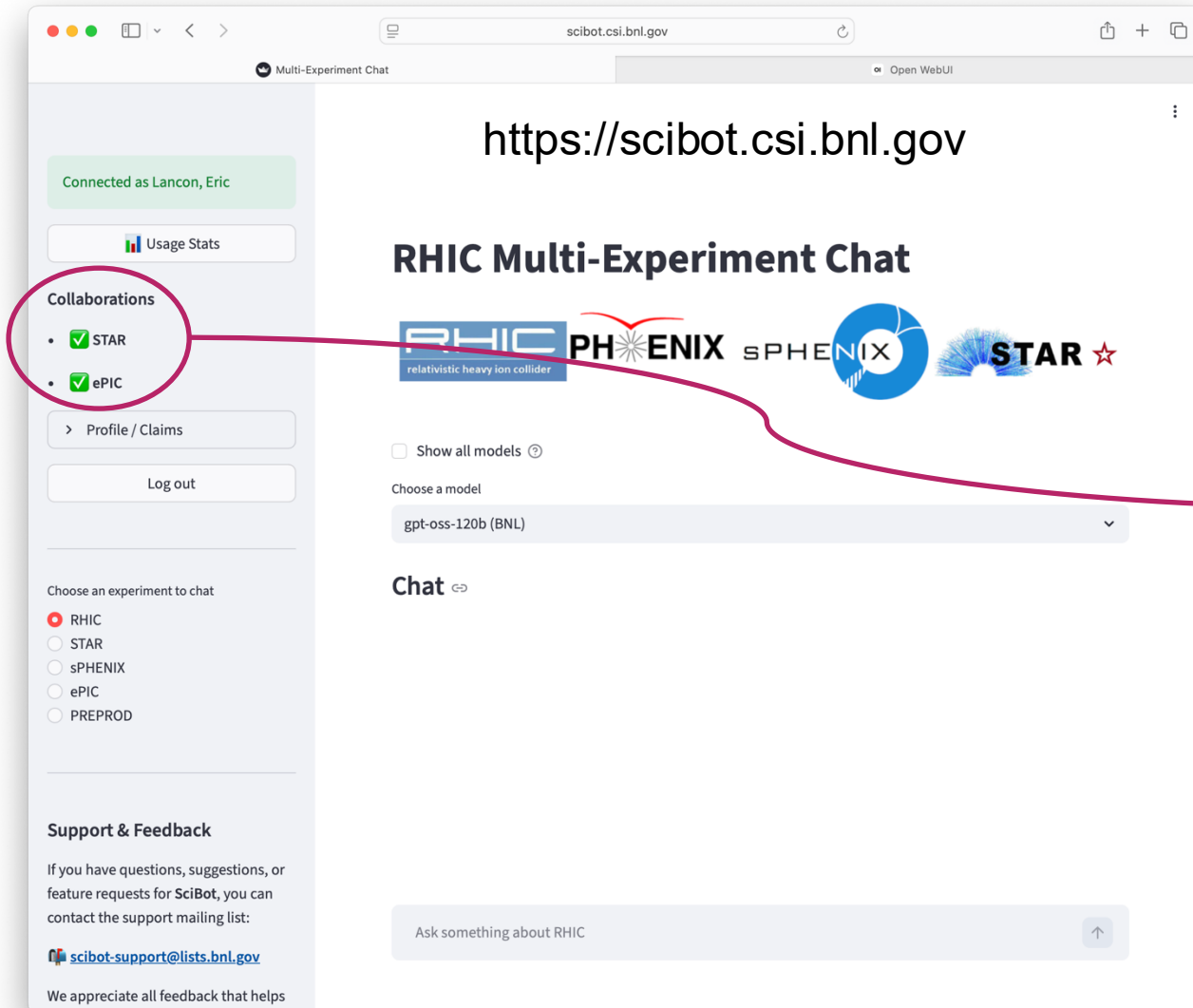
**Analysis = data + the knowledge and objects**  
needed to go from data to a physics result.

## A PHYSICS ANALYSIS IS MADE OF...



# SciBot update

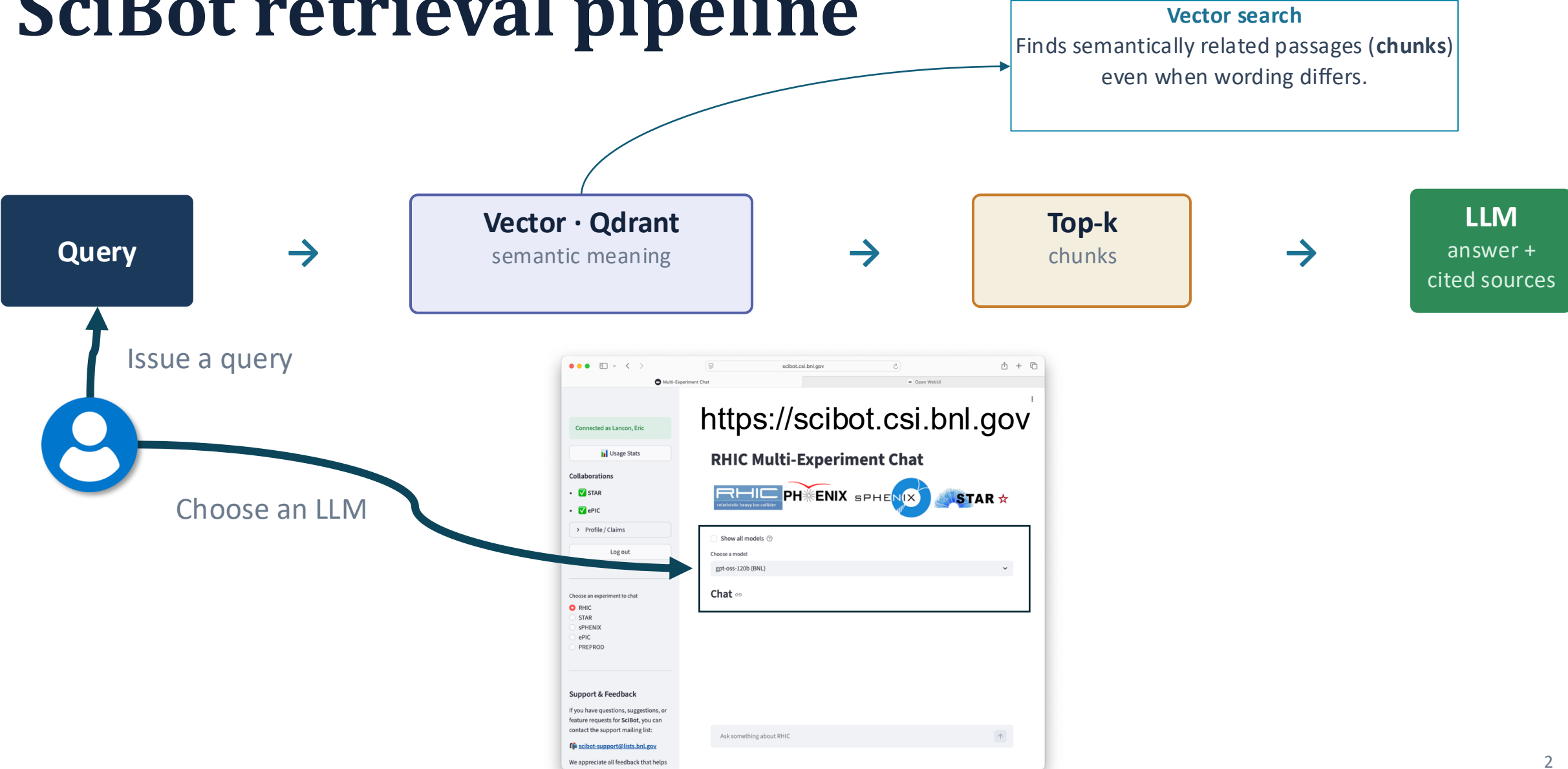
# SciBot



Interfaced with CoManage / PhoneBook

- **Federated identification:** no BNL-specific account required
- **Controlled authorization:** protects access to internal information

# SciBot retrieval pipeline



# Available LLMs

- BNL Inference Service
- DOE: ALCF inference Service(s)
  - Metis: Hot LLMs
  - Sophia: on demand

☐ Show all models ?

Choose a model

gpt-oss-120b (BNL)

gpt-oss-120b (BNL)

gemma-4-26b (BNL)

llama-4-scout-17b (BNL)

nemotron-3-super-120b (BNL)

gpt-oss-120b (DOE)

Mistral-Large-3-675B-Instruct-2512 (DOE)

A 'selected' list is proposed by default

- Not all models are shown



☒ Show all models ?

Choose a model

nemotron-3-ultra-550b-nvfp4

---- DOE Sophia ----

Meta-Llama-3.1-405B-Instruct

nemotron-3-super-120b

gpt-oss-120b (DOE)

Llama-3.3-70B-Instruct

Meta-Llama-3.1-70B-Instruct

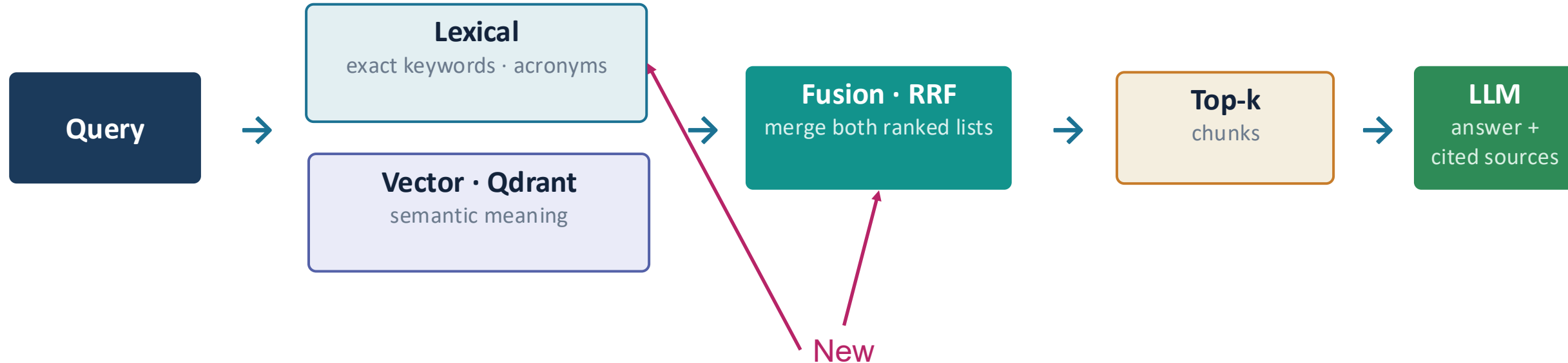
gemma-4-31B-it (DOE)

gemma-3-27b-it

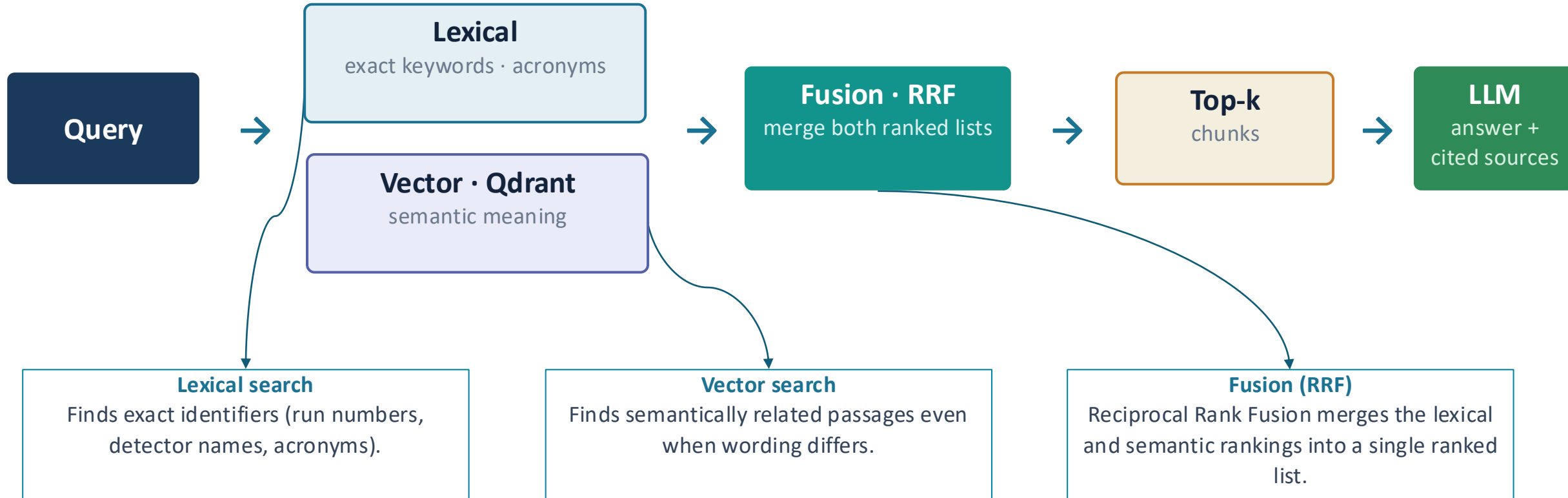
All available and relevant models are shown when option is selected

- Long list...

# Improved retrieval pipeline



# Improved retrieval pipeline



Example "STAR TPC gas mixture": the lexical search catches "TPC", the vector search catches the concept, and RRF merges them.

# Possible Scibot extensions

## BETTER RETRIEVAL

- 1 Reranking
- 2 Neighbor chunking
- 3 Metadata enrichment

Each extension addresses a different retrieval feature/issue.

## BETTER DOCUMENTS

- 4 Document quality

## BETTER ANSWERS

- 5 Grounding check
- 6 Recency

## NEW CAPABILITIES

- 7 Agentic retrieval
- 8 Graph RAG

# Retrieval improvements need to be quantified and validated

## 1. Measure retrieval performance

- Standard retrieval metrics to quantify improvements.
- Requires a set of expert-generated, verified Q&A pairs.
- Cannot rely on LLMs to generate grounded questions: LLM-generated questions have proven unreliable.
- Using multiple LLMs helps, but model disagreement can exceed the retrieval gain being measured.

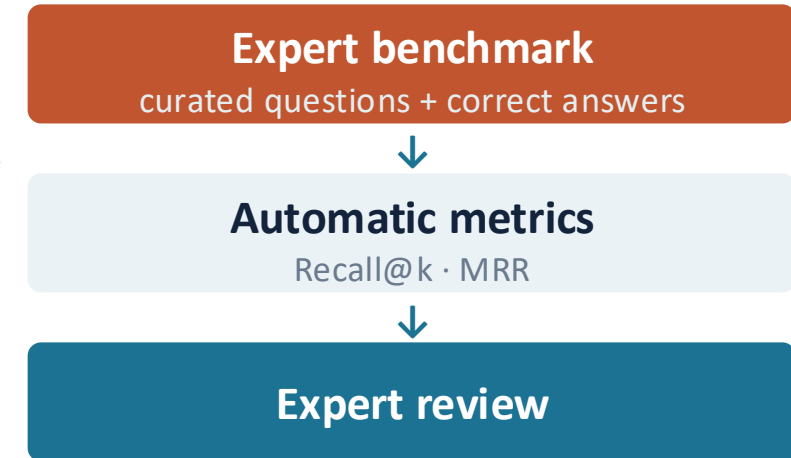
## 2. Validate the answers

- The final assessment also requires expert validation.
- LLM-as-a-judge has significant limitations:
  - Judges can contradict themselves: identical answer + quote → opposite verdicts
  - They may reward fluent, well-formatted answers even when incorrect

### Bottom line, we need:

- **O(100) expert-validated Q&A pairs**
- **Expert review of the evaluation results**

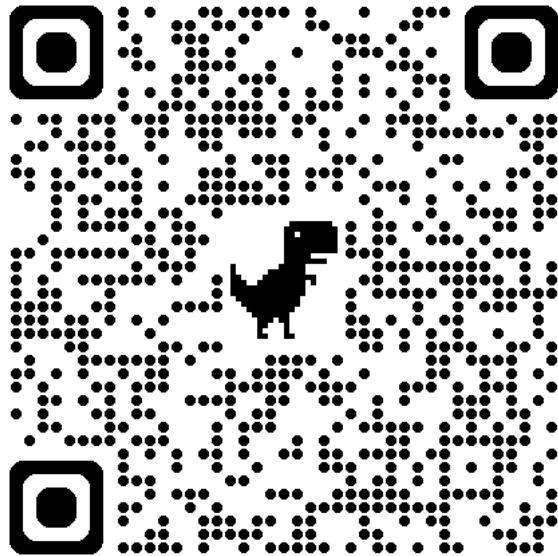
### THE EVALUATION LOOP



```
question : What safety factor is reported for the bolts under normal operation?
SciBot   : a minimum safety factor of 4
the note  : "The minimum safety factor for the normal loading condition was 4"
judge     : WRONG, the safety factor is 5.7  <- 5.7 is a DIFFERENT condition in the note
```

# Benchmarking questions

<https://forms.gle/6jimdRYouX1nWPgz6>



## Expert Benchmark Questions and Answers for SciBot

Help us build a high-quality benchmark for evaluating **SciBot** the AI assistant used at RHIC.

We are collecting expert-written questions together with their authoritative answers. These question/answer pairs will serve as a reference dataset for measuring retrieval accuracy, answer quality, and scientific correctness.

Please contribute questions that:

- have a unique, verifiable answer,
- can be answered from collaboration documentation or official sources,
- require scientific understanding rather than simple keyword lookup,
- reflect questions that collaborators commonly ask.

Each submission should include:

- the question,
- the expected answer,
- the authoritative source (URL, note, paper, manual, etc.),
- optional comments explaining why the question is important.

Expert-generated benchmark questions provide a reproducible way to compare SciBot implementations and to measure improvements over time.

[Sign in to Google](#) to save your progress. [Learn more](#)

\* Indicates required question

### Question \*

Write a question that an expert or collaborator might ask and make sure it has one verifiable answer.

Your answer

### Expected answer \*

# Next meeting(s)

Aug. 27 : Henry Klest (ANL) – AI assisted reanalysis of H1 data

Sep. 10 : Zhiwan XU (ANL) – AI-ready data for ePIC

Sep. 24: Ben Nachman (SLAC) – Reanalysis of SLD data – TBD

