



SciBot: Your "AI Librarian" for Navigating RHIC's Fragmented Institutional Knowledge

2026 RHIC/AGS ANNUAL USERS' MEETING

Ankush Kanuganti
for the BNL/DAP Team (M. Atif, Z. Dong, V. Garonne, J. Lauret, E. Lancon, A. Prozorov)



A 25-year scientific legacy faces a looming knowledge cliff

THE LEGACY

25 Years of Operations
(PHENIX, STAR, sPHENIX)



OPS_SPAN: 2000-2026

1 Exabyte of Irreplicable Data



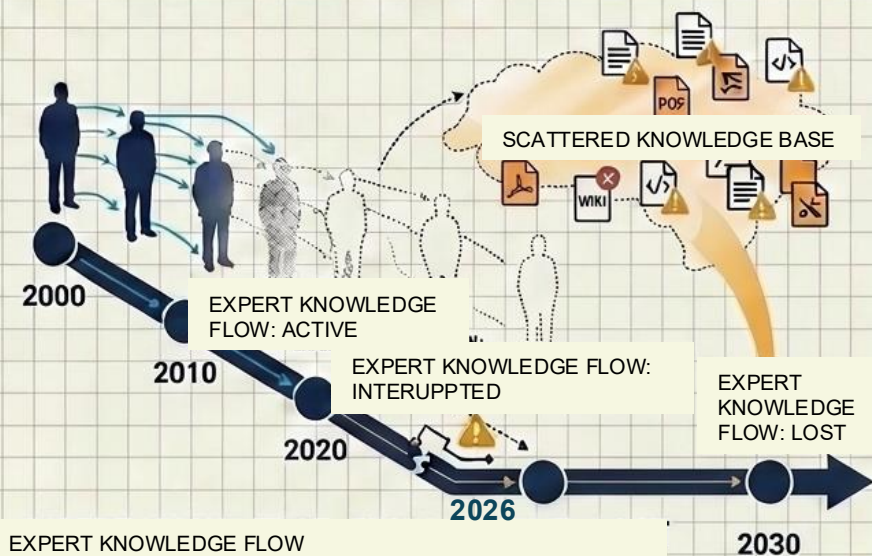
DATA_VOL: 1EB+

600+ Publications



PUB_COUNT: >680

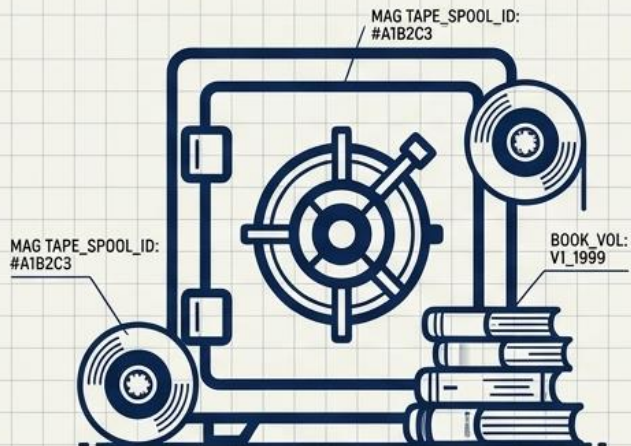
THE RISK



Expert detector and software knowledge is disappearing. Preservation must act before operations fully cease in 2026, with very limited knowledge by 2030.

The paradigm shift: From storing bits to preserving usability

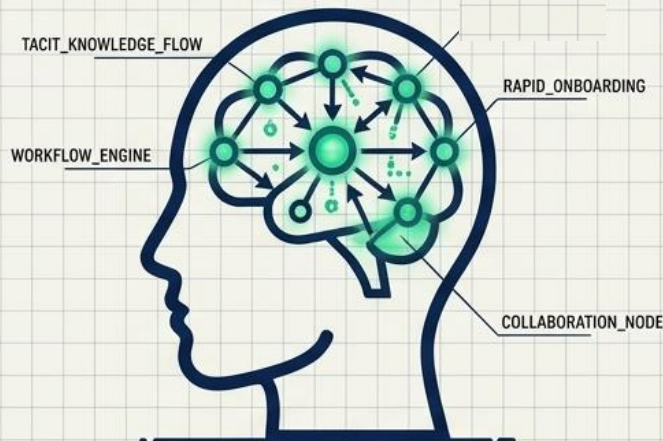
Bit Preservation



Storing files securely. Ensures data is safe, redundant, and verified.

Metaphor: A locked vault of books.

Functional Preservation

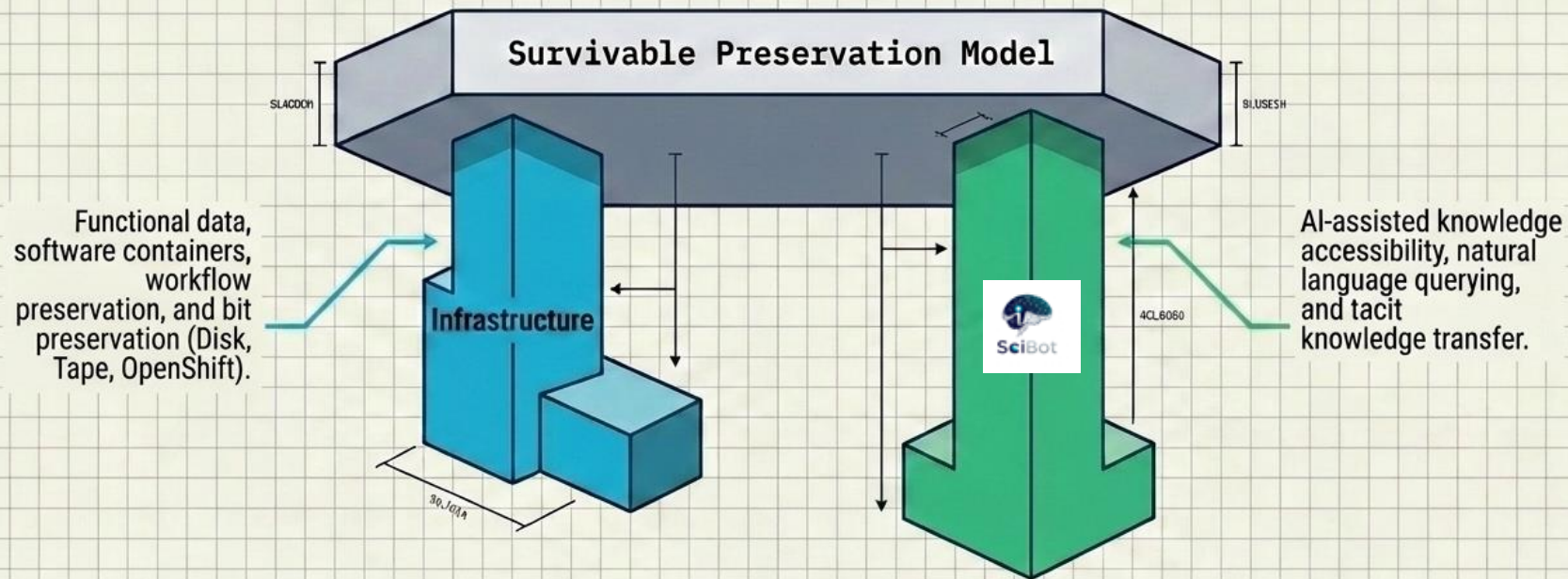


Maintaining usability. Capturing tacit knowledge, sustaining workflows, and enabling rapid onboarding for new collaborators.

Metaphor: The expert librarian who knows exactly how to fetch, read, and interpret the books.

The goal is not just to store data. The goal is to preserve the ability to use it.

A holistic framework for scientific continuity



None of these pillars are sufficient alone.

Without SciBot, preserved data becomes difficult to access.

Why generic commercial AI fails at institutional knowledge retrieval

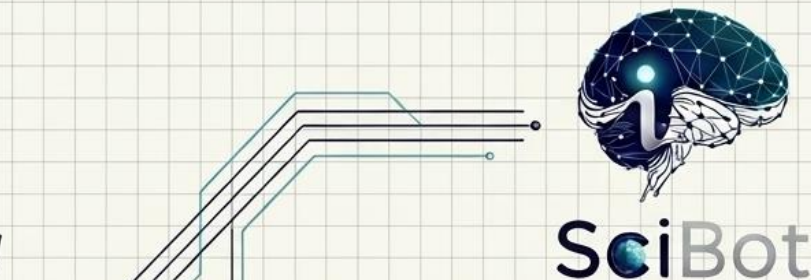
Dimensions of Evaluation	Generic Commercial API (e.g., ChatGPT)	SciBot (Local Domain RAG)
Access to Restricted Data & Collaboration-Scoped Control	❌ No access. Impossible to enforce role-based scopes.	✅ Direct secure access with built-in role-based filtering.
Data Privacy & Sovereignty	❌ Fails internal routing policies	✅ 100% local deployment guarantees sovereignty

Takeaway: Commercial models cannot enforce STAR-only vs. sPHENIX-only visibility. Local RAG models are an absolute requirement for sensitive scientific environments. Some technical documentation will never be released or made accessible to public.

SciBot: The Engine for Functional Preservation



25 years of fragmented institutional knowledge (Webpages, ArXiv, Theses, Analysis Notes, Document Repositories, etc.).



Cross-document semantic search queries to retrieve relevant information across multiple documents, rather than relying on exact keyword matches.



Natural language queries

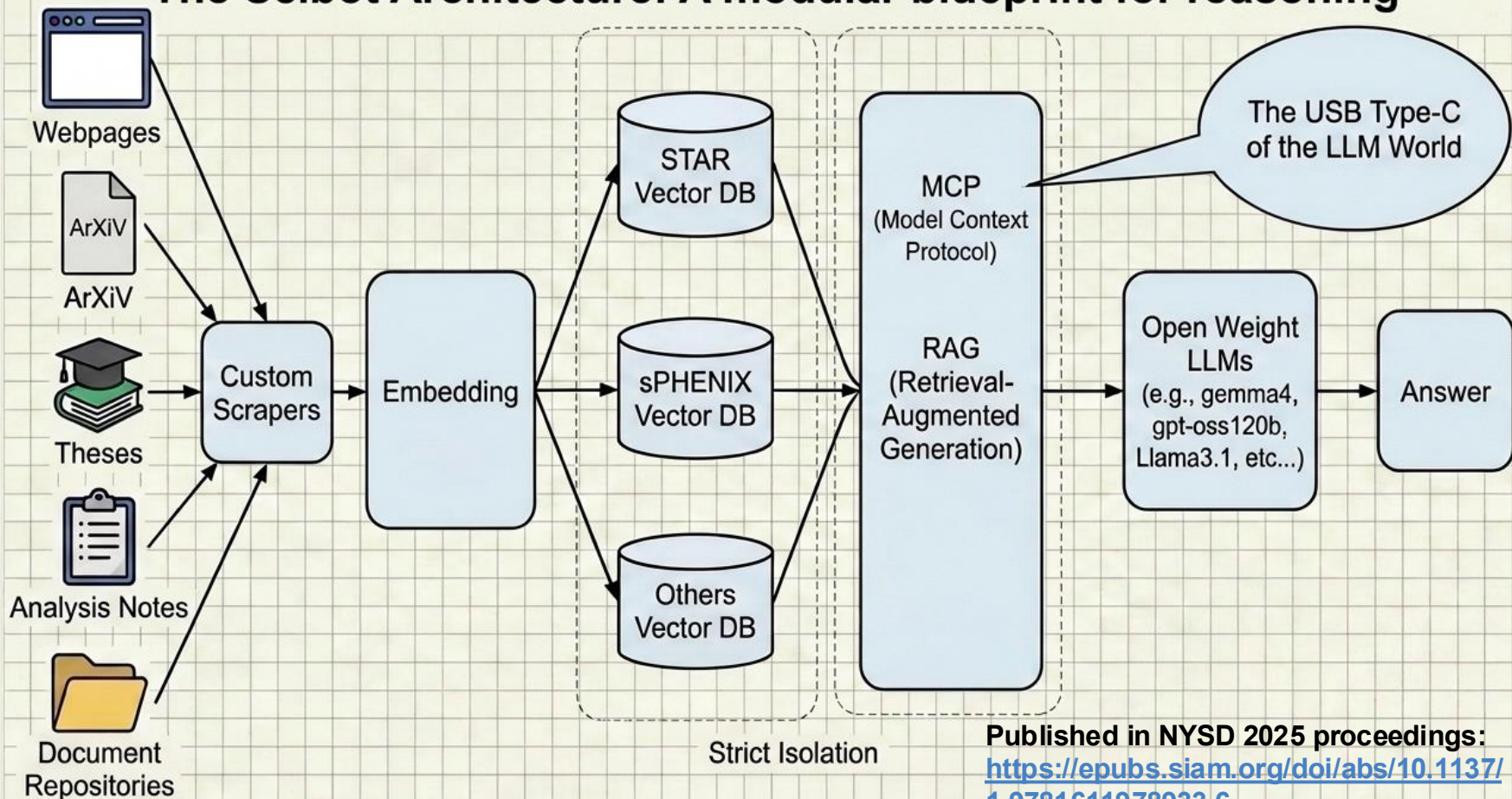


Model-agnostic RAG (**Retrieval-Augmented generation**) architecture combines user query with relevant chunks from internal sources for context-aware answers



Local deployment for data sovereignty

The Scibot Architecture: A modular blueprint for reasoning



Published in NYSD 2025 proceedings:
<https://epubs.siam.org/doi/abs/10.1137/1.9781611978933.6>

Current status of data sources ingested into the Scibot

Instance		RHIC	sPHENIX	STAR
www.bnl.gov	~ 15,000 webpages	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
www.star.bnl.gov	~ 6,700 webpages	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
www.sphenix.bnl.gov	~ 400 webpages	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
www.phenix.bnl.gov	~ 200 webpages	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Event Management pages (Indico)	Events up to Jan 30, 2026	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Arxiv	~ 820	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
Theses	~ 280	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
Analysis notes	~ 420	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
Mattermost Chats	~ 117 channels		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Zenodo	~ 740 PDFs	<input checked="" type="checkbox"/>		

The Front-End Experience: Unified access, tailored boundaries

- Ready for you to try out @ <https://scibot.csi.bnl.gov/>
- Remote inference is provided by **ALCF**
- ePIC is also added

Guest View

Authentication

Guest mode

Authenticate

Support & Feedback

If you have questions, suggestions, or feature requests for SciBot, you can contact the support mailing list:

scibot-support@lists.bnl.gov

We appreciate all feedback that helps us improve the service.

Multi-Experiment Chat

RHC relativistic heavy ion collider PHENIX sPHENIX STAR

Choose a model

openai/gpt-oss-120b

---- remote ----

openai/gpt-oss-120b

google/gemma-4-31B-it

meta-llama/Meta-Llama-3.1-70B-instruct

google/gemma-3-27b-it

---- local ----

gpt-oss-120b

llama3:latest

Ask something about RHC

Authenticated View

Authentication

Connected as Ankush Kanuganti

Collaborations

- ✓ STAR

> Profile / Claims

Log out

Choose an experiment to chat

RHC

STAR

Support & Feedback

If you have questions, suggestions, or feature requests for SciBot, you can contact the support mailing list:

scibot-support@lists.bnl.gov

We appreciate all feedback that helps us improve the service.

STAR Chatbot

STAR

Choose a model

openai/gpt-oss-120b

---- remote ----

openai/gpt-oss-120b

google/gemma-4-31B-it

meta-llama/Meta-Llama-3.1-70B-instruct

google/gemma-3-27b-it

---- local ----

gpt-oss-120b

llama3:latest

A single interface aggregates decades of fragmented knowledge across RHC, PHENIX, sPHENIX, and STAR. What previously required weeks of manual archive searching is reduced to seconds of verified, cited retrieval.

Proprietary LLM vs Scibot

ChatGPT >

How can I analyze jets in 200 GeV Au+Au collisions from 2014 in a RHIC STAR analysis. Explain in 250 words, cite each claim, restrict to RHIC domain only and keep the response grounded with zero fabrication.

A typical jet analysis in 200 GeV Au+Au collisions from the 2014 run at the STAR Collaboration follows a well-established workflow grounded in RHIC-specific reconstruction and heavy-ion techniques.

First, events are selected using minimum-bias or high-tower triggers, with vertex and centrality cuts derived from charged-particle multiplicity in the Time Projection Chamber and forward detectors (standard STAR event selection procedures). Charged tracks (TPC) and neutral energy (Barrel Electromagnetic Calorimeter) are combined to form particle candidates, ensuring full jet reconstruction within STAR acceptance ($|\eta| \leq 1$).

Jets are reconstructed using sequential recombination algorithms such as anti-k, via FastJet, typically with resolution parameter $R = 0.2-0.4$. This choice balances capture of jet energy against large underlying-event background in Au+Au (standard in RHIC jet analyses). The dominant complication is the heavy-ion background: a large, fluctuating soft particle density. This is corrected using area-based subtraction ($p \times A$), where p is the median background density estimated event-by-event (as established in RHIC heavy-ion jet measurements).

Detector effects and residual background fluctuations are handled through embedding: simulated p+p jets are embedded into real Au+Au events and reconstructed, allowing response matrices to be built. These matrices are then unfolded (e.g., iterative Bayesian methods) to obtain corrected jet spectra (standard STAR unfolding workflow).

Detector effects and residual background fluctuations are handled through embedding: simulated p+p jets are embedded into real Au+Au events and reconstructed, allowing response matrices to be built. These matrices are then unfolded (e.g., iterative Bayesian methods) to obtain corrected jet spectra (standard STAR unfolding workflow).

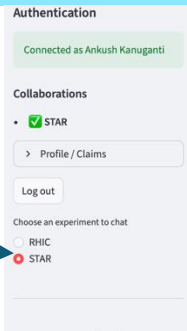
Physics observables include inclusive jet spectra, nuclear modification factor R_{AA} (comparison to p+p reference at 200 GeV), and jet substructure or hadron-jet correlations to probe quenching. Prior STAR measurements have shown strong jet suppression and broadening consistent with parton energy loss in the quark-gluon plasma at Relativistic Heavy Ion Collider energies.

Finally, systematic uncertainties (tracking efficiency, energy scale, background modeling) are quantified through variations in cuts and embedding configurations, following STAR analysis conventions.

Selecting authorized vector DB

Specific dataset selection

Citing internal documents



STAR Chatbot



Choose a model

gpt-oss-120b

Chat

How can I analyze jets in 200 GeV Au+Au collisions from 2014 in a RHIC STAR analysis

Steps to analyze 200 GeV Au+Au jets from the 2014 STAR run

- Data set & event selection
 - Use the Au+Au $\sqrt{s_{NN}}=200$ GeV data recorded in 2014 by STAR [4].
 - Select events with a high-energy Barrel Electromagnetic Calorimeter (BEMC) tower ($E_T > 4.3$ GeV) to define the *signal jet* sample; employ minimum-bias events for background estimation [4].
 - Apply the standard STAR vertex cut $|\nu_z| < 30$ cm for inclusive jet analyses ($|\nu_z| < 70$ cm can be used for $y_{dir} + jet$ studies) [7].
- Detector inputs
 - Charged tracks from the Time Projection Chamber (TPC) (full azimuth, $|\eta| < 1$).
 - Neutral energy from BEMC towers, with double-counting corrections for electrons and hadronic deposits [6].
- Jet reconstruction
 - Run the anti-k_T algorithm (FastJet) with radius parameter $R=0.4$ (or $R=0.3$ for tighter jets) [4, 5].
 - Include both charged tracks ($p_T \geq 200$ MeV/c) and calibrated BEMC towers in the clustering [5].
- Background subtraction
 - Compute the event-wise median background density p and correct each jet's raw transverse momentum: $p_{T,corr} = p_{T,rec} - p \cdot A$ ($A = jet\ area$) [5].
 - Use a mixed-event technique to model the combinatorial jet background and subtract it statistically from the measured spectra [4].
- Bias mitigation
 - To obtain a trigger-bias-free jet sample, study recoil jets opposite a high- p_T hadron or photon trigger; the recoil jet population is largely unbiased [4].

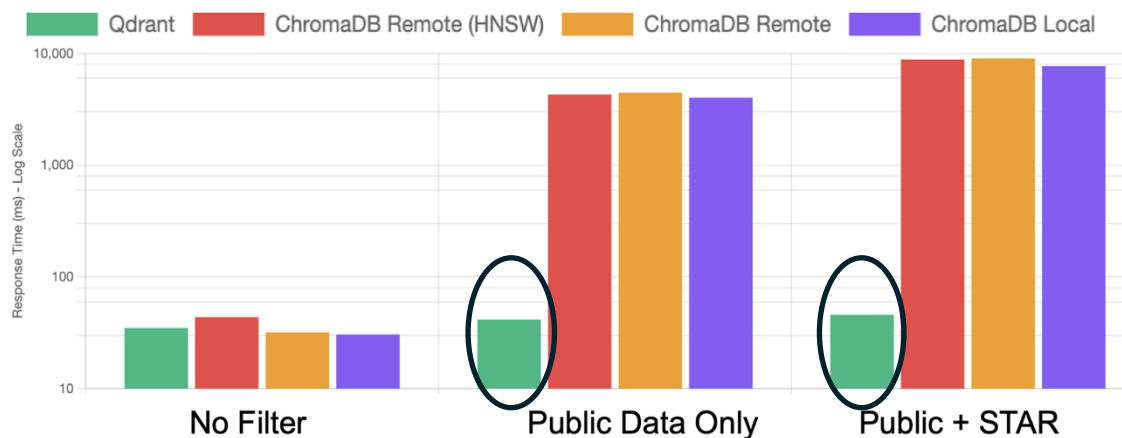
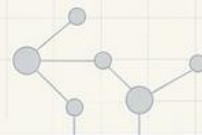
Federated identity: Enforcing strict 'Need to Know' boundaries



Non-members never receive restricted chunks. Access control is enforced at the database retrieval level, ensuring the LLM physically cannot leak cross-collaboration secrets.

```
{
  "$or": [
    {"visibility": "public"},
    {"collaboration": {
      "$in": user.collaborations}}
  ]
}
```

The Speed Imperative: Pivoting to Qdrant



Context Panel

Why the Pivot?

Requirement:

Users demand fast inference.

Challenge:

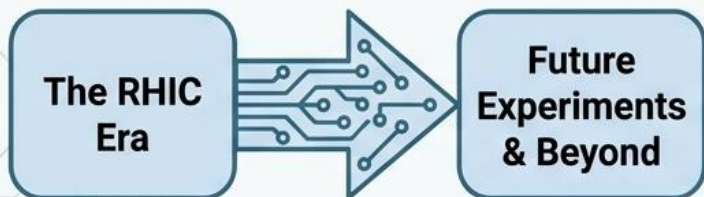
Segregation of information (STAR vs. sPHENIX) requires heavily filtered queries. ChromaDB was too slow under filtered queries at scale.

Solution:

Qdrant ensures real-time user experiences even when searching through deeply restricted metadata schemas.



Conclusion: A transferable blueprint for central knowledge discovery



Takeaway



1. Mitigating Expert Knowledge Risk & Streamlining Onboarding: SciBot helps mitigate expert knowledge risk and assists new physicists in navigating various internal techniques.

2. SciBot: The Librarian for RHIC Experiments: Serves as the central knowledge discovery tool, ensuring reproducibility and providing AI-ready data access for all RHIC experiments.

3. Adaptable to Any Experiment: Easily adaptable for experiments with scattered, diverse data sources due to flexible ingestion pipelines and rigorous testing.

4. Integrating with other BNL services: Based on the success of the scibot's performance, BNL administration is working to integrate services like SBMS into this system.



Glossary

Retrieval-Augmented Generation (RAG): This core architecture enables the bot to answer questions by first retrieving semantically relevant chunks from a vast database of RHIC documents and then using an LLM to generate a response based on that specific context. This method significantly reduces hallucinations, ensures domain-specific accuracy, and provides traceable citations for every answer.

Model Context Protocol (MCP): Described as the "USB Type-C" of the LLM world, MCP provides a modular reasoning pipeline. It standardizes the flow from retrieval and summarization to inference and evaluation, allowing the system to work flexibly with various local GPUs or external APIs.

Vector Databases (ChromaDB and Qdrant): These databases store document embeddings as vectors to allow for high-speed semantic searches. While the project initially used ChromaDB, it transitioned to Qdrant because it provides significantly faster response times when handling complex filtered queries required for data segregation.

Federated Identity and Access Control: To protect sensitive or restricted experimental data, the bot utilizes a security stack including PhoneBook, COmanage, and CILogon. This ensures that only authenticated members of specific collaborations (like STAR or sPHENIX) can access their respective internal documents, while guest users are limited to public information.

Custom Scraper Framework: This framework is designed for multi-format parsing, converting HTML, PDFs, Word documents, and PowerPoint presentations into text while preserving critical metadata like URLs and timestamps. It includes smart filtered crawling and rate-limiting to prevent server overloads during data ingestion.

Inference Engines: The system tests several engines to optimize performance, including vLLM, LlamaCPP, and Ollama. Benchmarking shows that vLLM scales best for multi-GPU setups, while LlamaCPP is efficient for single-GPU operations.

Engineering the prompt: Hard guardrails for scientific fidelity

Domain Hard Boundary

Restricts answers exclusively to the RHIC domain.

Grounded Only

Must use retrieved database context. Zero external knowledge fabrication.

To analyze jets in 200 GeV Au+Au collisions from 2014, you can follow the steps outlined below:

1. **Event selection:** Select events that have a high-energy trigger tower in the Barrel Electromagnetic Calorimeter (BEMC) and at least one charged track with $p_T > 0.2$ GeV/c in the Time Projection Chamber (TPC). The trigger towers should be within the pseudorapidity range of $|\eta| < 1.0$.
2. **Jet reconstruction:** Use the anti- k_r jet finding algorithm with a resolution parameter $R = 0.4$ to reconstruct jets from the charged tracks in the TPC. The jets should have $p_T > 2$ GeV/c and be within the pseudorapidity range of $|\eta| < 1 - R$.
3. **Background subtraction:** Subtract the underlying event background using a method based on active jet areas [1, 2]. This is important to ensure that the jets are not contaminated by non-jetlike particles.
4. **Jet shape analysis:** Study the differential jet shapes as a function of r , where r is the radial distance from the jet axis. You can use the mixed-event technique to estimate background contributions and subtract them from each p_T^{assoc} range [3].
5. **Systematic errors:** Consider systematic errors related to the jet finding procedure, such as the choice of jet resolution parameter R . You can study the effect of changing R on the jet shapes.
6. **Comparison with p+p collisions:** Compare your results with those from 200 GeV p+p collisions at similar center-of-mass energies. This will help you understand whether there are any significant modifications to the jets in Au+Au collisions.

References:

- [1] STAR Collaboration, Phys. Rev. C 92 (2015) 054906
- [2] STAR Collaboration, Phys. Rev. C 94 (2016) 024908
- [3] STAR Collaboration, Phys. Rev. C 95 (2017) 034903

Format Constraints

Strictly limited to 200 words, utilizing bullet points for rapid scannability.

Citation Discipline

Every factual claim must tie to a numbered, verifiable source document identifier.

Hmm, I'm not sure.

Graceful Fallback

The 'I Don't Know' rule.

Proprietary LLM vs SciBot

ChatGPT > ...

Can you point out to the references or citations?

Here are authoritative RHIC/STAR references that correspond directly to each step I mentioned. I'm keeping this strictly within the RHIC domain and avoiding unsupported claims.

Experimental setup, detectors, and event selection

- STAR Collaboration detector overview:
 - K.H. Ackermann et al., *Nucl. Instrum. Meth. A* 499 (2003) 624 — STAR detector description (TPC, BEMC, triggers).
- Time Projection Chamber:
 - M. Anderson et al., *Nucl. Instrum. Meth. A* 499 (2003) 659.
- Barrel Electromagnetic Calorimeter:
 - M. Beddo et al., *Nucl. Instrum. Meth. A* 499 (2003) 725.

Jet reconstruction and algorithms

- FastJet:
 - M. Cacciari, G.P. Salam, G. Soyez, *Eur. Phys. J. C* 72 (2012) 1896 (FastJet manual).
- Anti-k_s algorithm: M. Cacciari, G.P. Salam, G. Soyez, *JHEP* 04 (2008) 063.

Heavy-ion background subtraction (RHIC context)

- Area-based subtraction ($p \times A$):
 - M. Cacciari, G.P. Salam, *Phys. Lett. B* 659 (2008) 119.
- STAR implementation in Au+Au:
 - L. Adamczyk et al. (STAR), *Phys. Rev. Lett.* 115, 092002 (2015) — inclusive jet measurements at 200 GeV Au+Au.

Internal Resources

Authentication

Guest mode

Authenticate

Support & Feedback

If you have questions, suggestions, or feature requests for SciBot, you can contact the support mailing list:

scibot-support@lists.bnl.gov

We appreciate all feedback that helps us improve the service.

References

- [1] <https://arxiv.org/abs/2002.06217v2>
- [2] https://drupal.star.bnl.gov/STAR/files/phd_thesis_rusnak.pdf
- [3] <https://arxiv.org/abs/2309.00156v3>
- [4] <https://arxiv.org/abs/0907.4788v3>
- [5] https://indico.bnl.gov/event/15148/attachments/40846/68609/STAR_BUR_Runs23_25_2022%201.pdf
- [6] <https://drupal.star.bnl.gov/STAR/starnotes/public/sn0399>
- [7] https://indico.bnl.gov/event/20331/attachments/49219/83916/STAR_BUR_Runs24_25_2023.pdf
- [8] https://indico.bnl.gov/event/5039/contributions/26250/attachments/21609/29603/jet_substructure_star_santafeJetWorkshop_UCLA_jan2019_raghav_v3.pdf

> Details

Multi-Experiment Chat



Choose a model

gpt-oss-120b

---- local ----

gpt-oss-120b

llama3:latest